

# Enzyklopädie der Psychologie

# ENZYKLOPÄDIE DER PSYCHOLOGIE

In Verbindung mit der  
Deutschen Gesellschaft für Psychologie

herausgegeben von

Prof. Dr. Carl F. Graumann, Heidelberg

Prof. Dr. Theo Herrmann, Mannheim

Prof. Dr. Hans Hörmann, Bochum

Prof. Dr. Martin Irle, Mannheim

Prof. Dr. Dr. h.c. Hans Thomae, Bonn

Prof. Dr. Franz E. Weinert, München

Themenbereich B

Methodologie und Methoden

Serie I

Forschungsmethoden der Psychologie

Band 5

Hypothesenprüfung



Verlag für Psychologie · Dr. C. J. Hogrefe  
Göttingen · Toronto · Zürich

# Hypothesenprüfung

Herausgegeben von

Prof. Dr. Jürgen Bredenkamp, Trier  
und Prof. Dr. Hubert Feger, Hamburg



Verlag für Psychologie · Dr. C. J. Hogrefe  
Göttingen · Toronto · Zürich

© by Verlag für Psychologie Dr. C.J. Hogrefe, Göttingen 1983  
Alle Rechte, insbesondere das der Übersetzung in fremde Sprachen, vorbehalten.

Gesamtherstellung: Allgäuer Zeitungsverlag GmbH, 8960 Kempten (Allgäu)  
Printed in Germany

ISBN 3 8017 0515 3

# Autorenverzeichnis

*Prof. Dr. Jürgen Bredenkamp*

Fachbereich I - Psychologie  
der Universität Trier  
Schneidershof  
D - 5500 Trier

*Dr. Willi Nagl*

Universität Konstanz  
Turnerstraße 6  
D - 7750 Konstanz

*Dr. Willi Hager*

Institut für Psychologie  
der Universität Göttingen  
Goßlerstraße 12  
D - 3400 Göttingen

*Dr. Hans Ueckert*

Psychologisches Institut II  
der Universität Hamburg  
Von-Melle-Park 5  
D - 2000 Hamburg 13

*Prof. Dr. Claus Möbus*

Fachbereich Math./Informatik  
der Universität Oldenburg  
Ammerländer Heerstraße 67-99  
D - 2900 Oldenburg

*Prof. Dr. Dirk Wendt*

Institut für Psychologie  
der Christian-Albrechts-Universität  
Olshausenstraße 40/60  
D - 2300 Kiel

*Dr. Rainer Westermann*

Institut für Psychologie  
der Universität Göttingen  
Goßlerstraße 12  
D - 3400 Göttingen

# Inhaltsverzeichnis

## 1. Kapitel: Übersicht. Von Jürgen Bredenkamp

1. Zur Prüfung sog. Kausalhypothesen . . . . .	1
2. Zum Problem der Validität des statistischen Schlusses . . . . .	14
3. Dynamische Modelle . . . . .	18

## 2. Kapitel: Planung und Auswertung von Experimenten. Von Willi Hager und Rainer Westermann

Vorbemerkungen . . . . .	24
1. Einleitung. . . . .	26
1.1 Einige Begriffsbestimmungen . . . . .	26
1.2 Das Experiment als Methode zur Prüfung von Kausalaussagen . . . . .	27
1.3 Die Validität eines Experiments . . . . .	29
2. Variablenvalidität (VV) . . . . .	33
2.1 Mangelnde Eindeutigkeit der Zuordnung als Störfaktor (VV) . . . . .	34
2.2 Mangelnde konzeptuelle Replikation als Störfaktor (VV) . . . . .	35
2.3 Mangelnde Entsprechung im Variationsbereich von theoretischen und empirischen Variablen als Störfaktor (VV) . . . . .	37
2.4 Zu geringes Skalenniveau als Störfaktor (VV) . . . . .	38
2.5 Konfundierung von theoretischen Begriffen als Störfaktor (VV) . . . . .	42
2.6 Zusammenfassung . . . . .	46
3. Interne Validität . . . . .	46
3.1 Variation personaler und situationaler Merkmale als Störfaktoren (IV) . . . . .	47

3.1.1 Variation situationaler Merkmale . . . . .	47
3.1.2 Variation personaler Merkmale . . . . .	48
3.2 Störfaktoren (IV) bei Meßwiederholung . . . . .	48
3.3 Zur Kontrolle der Störfaktoren (IV) bei interindividueller Bedingungs- variation . . . . .	50
3.3.1 Konstanthaltung und Elimination . . . . .	51
3.3.2 Randomisierung . . . . .	52
3.3.3 Einführung eines Kontrollfaktors . . . . .	55
3.4 Zur Kontrolle der Störfaktoren (IV) bei intraindividueller Bedingungs- variation (Meßwiederholung) . . . . .	56
3.5 Versuchspläne mit interindividueller Bedingungsvariation und Vortest . . . .	58
3.6 Zur Definition des Experiments und anderer Untersuchungsmethoden . . .	59
4. Populations- und Situationsvalidität . . . . .	60
4.1 Populationsvalidität (PV) . . . . .	60
4.2 Situationsvalidität (SV) . . . . .	63
4.3 Zur Kontrolle der Störfaktoren (PV und SV) . . . . .	64
5. Beziehungen zwischen den Validitätsarten . . . . .	65
6. Statistische Validität . . . . .	67
7. Eine Strategie zur Entscheidung zwischen statistischen Hypothesen: Der Signifikanztest . . . . .	70
7.1 Überblick über verschiedene alternative Strategien . . . . .	70
7.2 Kurzer Abriß einiger Charakteristika von Signifikanztests . . . . .	73
7.3 Mögliche Fehler beim statistischen Testen . . . . .	76
7.3.1 Fehler unter Gültigkeit der Null-Hypothese (Fehler 1. Art) . . . . .	76
7.3.2 Fehler unter Gültigkeit der Alternativhypothese (Fehler 2. Art) . . . . .	78
7.4 Die Determinanten eines Signifikanztests . . . . .	81
7.4.1 Forschungs- und Publikationspraxis I: Signifikanzniveau und p-werte . . . . .	83
7.4.2 Forschungs- und Publikationspraxis II: Experimentelle Effekte und Teststärke . . . . .	85
7.4.3 Forschungs- und Publikationspraxis III: Entwicklung einer vorläufigen Zielvorstellung . . . . .	86
7.5 Arten statistischer Hypothesen und ihre Prüfung . . . . .	88
7.5.1 Gerichtete und ungerichtete Hypothesen und ihre Prüfung . . . . .	88
7.5.2 Parametrische und nicht-parametrische Hypothesen und ihre Prüfung	89
7.5.3 Zur Wahl zwischen parametrischen und nicht-parametrischen Verfahren . . . . .	91

7.5.3.1 Auswertung von Häufigkeitsdaten (Nominal-Niveau) . . . . .	92
7.5.3.2 Auswertung von Rangdaten (Ordinal-Niveau) . . . . .	93
7.5.3.3 Auswertung von Intervalldaten (Intervall-Niveau) . . . . .	94
7.5.4 Zur Frage der relativen Effizienz . . . . .	95
7.6 Zusammenfassung . . . . .	96
8. Störfaktoren der statistischenvalidität und ihre Ausschaltung . . . . .	97
8.1 <i>Falsche statistische Hypothesen und Verfahren</i> . . . . .	97
8.1.1 Die wichtigsten Beziehungen zwischen psychologischen und statistischen Hypothesen . . . . .	97
8.1.2 Falsche Umsetzung der wissenschaftlichen in eine statistische Hypothese als Störfaktor (StatV) . . . . .	101
8.1.3 Falsche Auswahl der zu prüfenden statistischen Hypothese . . . . .	102
8.1.4 Falsche statistische Analyse . . . . .	102
8.2 <i>Verletzung der Annahmen bei statistischen Tests als Störfaktor (StatV)</i> . . . .	103
8.2.1 Das Allgemeine Lineare Modell (ALM) und die Annahmen . . . . .	103
8.2.2 Additivität . . . . .	106
8.2.3 Normalverteilung der Modellresiduen (Fehler) . . . . .	108
8.2.4 Homogenität der Fehlervarianzen in den Populationen . . . . .	111
8.2.4.1 Zur Frage des Prüfverfahrens bei Varianzheterogenität . . . . .	113
8.2.4.2 Zur Bedeutung von Transformationen . . . . .	114
8.2.5 Unabhängigkeit der Fehlerterme . . . . .	115
8.2.5.1 Zur Residuenanalyse; Ausreißerwerte . . . . .	116
8.2.6 Problem der Zufallsstichproben . . . . .	119
8.3 <i>Kumulierung der Wahrscheinlichkeiten für Fehler erster und zweiter Art</i> . . . .	120
8.3.1 Multiple Mittelwertsvergleiche . . . . .	123
8.3.2 Monotone Trendhypothesen . . . . .	126
8.4 <i>Mangelnde Präzision</i> . . . . .	127
8.4.1 Parallelisierung als Kontrolltechnik (StatV) . . . . .	128
8.4.2 Kovarianzanalyse als Kontrolltechnik (StatV) . . . . .	130
8.4.3 Homogenisierung als Kontrolltechnik (StatV) . . . . .	131
8.4.4 Konstanthaltung und Elimination als Kontrolltechniken (StatV) . . . .	131
8.4.5 Eingenistete Faktoren als Kontrolltechnik (StatV) . . . . .	132
8.4.6 Wiederholte Messungen als Kontrolltechnik (StatV) . . . . .	133
8.4.6.1 Analyse von Zeitreihen und Veränderungsmessungen . . . . .	136
8.4.6.2 Univariate und multivariate Analysen . . . . .	137
8.4.6.2.1 Exakte und approximative univariate Tests . . . . .	137
8.4.6.2.2 Multivariate Tests . . . . .	142
8.4.6.2.3 Nicht-parametrische Tests . . . . .	143
8.4.7 Zur Beziehung zwischen der Präzision und den anderen Aspekten der experimentellen Validität . . . . .	143



8.5	<i>Falsche Analyse und Interpretation statistischer Interaktionen</i>	144
8.5.1	Das Konzept der statistischen Interaktion	145
8.5.2	Definition verschiedener Typen der Interaktion	148
8.5.2.1	Disordinale Interaktion	149
8.5.2.2	Ordinale Interaktion	150
8.5.2.3	Zur graphischen Darstellung von Interaktionen	152
8.5.3	Ein Verfahren zur Unterscheidung zwischen den Interaktionstypen	154
8.6	<i>Zusammenfassung</i>	155
9.	Maße der statistischen Assoziation: Die experimentellen Effekte	157
9.1	<i>Einleitung</i>	157
9.2	<i>Experimentelle Effekte und praktische Bedeutsamkeit</i>	158
9.3	<i>Experimentelle Effekte bei parametrischen Hypothesen</i>	158
9.3.1	Maße der Nicht-Zentralität: $\lambda$ , $\varphi^2$ und $f^2$	159
9.3.2	Korrelationskoeffizienten und -quotienten	160
9.3.2.1	Populationsmaße: $\eta^2$ , $\omega^2$ und $R^2_{YX}$	161
9.3.2.2	Stichprobenmaße: $\hat{R}^2_{YX}$ , $E^2$ , UI	163
9.3.2.3	Korrekturformeln für $\hat{R}^2_{YX}$	164
9.4	<i>Experimentelle Effekte bei nicht-parametrischen Hypothesen</i>	166
9.4.1	Experimentelle Effekte bei ordinalen Daten	166
9.4.2	Experimentelle Effekte bei nominalen Daten	167
9.5	<i>Zur Kritik der Maße der statistischen Assoziation</i>	168
9.6	<i>Zusammenfassung</i>	169
10.	Bestimmung des Stichprobenumfanges	170
10.1	<i>Überblick</i>	170
10.2	<i>Allgemeine Prinzipien der Stichprobengrößenbestimmung</i>	172
10.3	<i>Bestimmung des Stichprobenumfanges bei univariaten Varianz- und Regressionsanalysen</i>	174
10.3.1	Bei Kenntnis der Populationsvarianz $\sigma_e^2$ („Klassischer Ansatz“)	174
10.3.2	Bei prä-experimenteller Schätzung der Varianz $\sigma_e^2$ („Two-Stage-Sampling“-Verfahren nach Stein und Rodger)	175
10.3.3	Ohne Kenntnis der Populationsvarianz $\sigma_e^2$	177
10.3.3.1	Festlegung von $\varphi^2$	177
10.3.3.2	Festlegung von $f^2$ oder $R^2_{YX}$ (Verfahren nach Cohen)	177
10.4	<i>Hinweise zur Stichprobenumfangsbestimmung bei weiteren Gruppen von parametrischen Testverfahren</i>	180
10.4.1	Varianzanalyse mit zufälligen und gemischten Effekten	180
10.4.2	Nicht-orthogonale Varianzanalysen	181
10.4.3	Multivariate Varianz- und Regressionsanalysen	182

<i>10.5 Hinweise zur Stichprobenumfangsbestimmung bei nicht-parametrischen Verfahren . . . . .</i>	183
10.5.1 Nominale Daten . . . . .	183
10.5.2 Ordinale Daten . . . . .	183
<i>10.6 Abschließende Bemerkungen zur Stichprobengrößenbestimmung . . . . .</i>	184
 11. Eine Strategie zur Entscheidung über wissenschaftliche Hypothesen mittels Signifikanztests . . . . .	185
<i>11.1 Stadium der Planung des Experiments . . . . .</i>	185
11.1.1 Überblick . . . . .	185
11.1.2 Zur Festlegung der beiden Fehlerwahrscheinlichkeiten . . . . .	186
11.1.3 Zur Festlegung des experimentellen Mindesteffektes EEM . . . . .	188
11.1.4 Zur Frage der Willkür bei der Planung von Experimenten . . . . .	188
<i>11.2 Stadium der Entscheidung über die Kausalhypothese . . . . .</i>	188

### 3. Kapitel: Messung, Analyse und Prognose von Veränderungen. Von Claus Möbus und Willi Nagl

1. Einleitung. . . . .	239
2. Univariate Zeitreihenanalyse . . . . .	243
2.1 Integrierte Prozesse der Ordnung $d$ : $ARIMA(0,d,0)$ -Modelle . . . . .	246
2.2 Autoregressive Prozesse der Ordnung $p$ : $ARIMA(p,0,0)$ und $ARIMA(p,d,0)$ -Modelle . . . . .	248
2.3 Moving-average Prozesse der Ordnung $q$ : $ARIMA(0,0,q)$ -Modelle . . . . .	251
2.4 Das allgemeine $ARIMA(p,d,q)$ -Modell . . . . .	253
2.5 Autokorrelations- und partielle Autokorrelationsfunktion . . . . .	256
2.6 Saisonale Einflüsse . . . . .	263
2.7 Modellidentifikation . . . . .	264
2.8 Multiple Zeitreihenanalyse: Transferfunktionsmodelle . . . . .	268
2.9 Multivariate Zeitreihenanalyse . . . . .	279
2.10 Multiple und multivariate Transfermodelle . . . . .	282
2.10.1 Multiple Transfermodelle . . . . .	282
2.10.2 Multivariate Transfermodelle . . . . .	283
3. Zeitreihenexperimente . . . . .	284
3.1 $N = 1$ -Experimente . . . . .	285
3.1.1 Verteilungsfreie Prüfmethode: Randomisierungs- bzw. Permutationstests . . . . .	287

3.1.2 Verteilungsgebundene Prüfverfahren: Lineares Modell . . . . .	293
3.1.3 Verteilungsgebundene Prüfverfahren: Interventionsanalyse mit dem Transfermodell von Box & Tiao (1975) . . . . .	304
3.2 $N > 1$ Quasiexperimentelle Zeitreihendesigns (univariater Fall für eine Gruppe) . . . . .	315
3.2.1 $N>T$ , $M=1$ , $G=1$ . . . . .	315
3.2.2 $G > 1$ Quasiexperimentelle Zeitreihendesigns bei mehreren Gruppen	327
3.2.3 $M > 1$ Quasiexperimentelle Zeitreihendesigns mit mehreren abhängi- gen Variablen . . . . .	327
4. Veränderungsmessung mit Hilfe von Differenzenwerten . . . . .	328
4.1 Korrelation zwischen Anfangswert und Differenzwert . . . . .	329
4.2 Schätzung individueller Veränderungswerte . . . . .	330
4.3 Der Differenz- bzw. Endwert in der Regressionsanalyse . . . . .	333
4.4 Kovarianz- bzw. Regressionsmodell bei zeitbezogenen Daten . . . . .	338
4.5 Reliabilität-Stabilität . . . . .	343
5. Wachstumskurven- und Varianzanalyse . . . . .	344
5.1 Der Eingruppenfall . . . . .	344
5.1.1 Der „wiederholte Messungen“-Ansatz ( $T \geq 2$ , $G = 1$ , $N > 1$ ) . . . . .	344
5.1.2 Zur Identifikation und Interpretation der Effektparameter . . . . .	349
5.2 Berücksichtigung von gruppenspezifischen Faktoren . . . . .	357
5.3 Schätzung des Modells. . . . .	364
5.4 Hypothesentests . . . . .	366
5.5 Mehrfachantwort (echt multivariate) -Analyse . . . . .	369
6. Pooling von „Querschnitt“ - mit „Zeitreihen“-Analyse . . . . .	370
6.1 Modellüberlegungen . . . . .	370
6.2 Schätzprobleme . . . . .	375
7. Strukturgleichungsmodelle . . . . .	375
7.1 Kovarianz- und korrelationsorientierte Analysen von Zeitreihen von Quer- schnitten: Stabilität von Konstrukten . . . . .	376
7.2 Wachstumskurvenanalyse als Strukturgleichungsmodell . . . . .	381
7.3 Erwartungswertorientierte Analysen von Zeitreihen von Querschnitten: Zeit- bezogene Hypothesen für diskrete Zeitpunkte . . . . .	384
7.4 Erwartungswertorientierte Analysen von Zeitreihen von Querschnitten: Schereneffekte bei Mittelwertsverläufen auf latenten Variablen . . . . .	390
8. Markoff-Modelle für qualitative Variable bei diskreter Zeit . . . . .	395
8.1 Markoffketten 1. Ordnung mit einer Variablen . . . . .	401
8.2 Markoffketten 2. Ordnung (1 Variable) . . . . .	407

8.3 Markoffketten mit mehreren Variablen . . . . .	409
8.4 Schätzung der Übergangswahrscheinlichkeiten . . . . .	410
8.4.1 bei Zeitinhomogenität . . . . .	411
8.4.2 bei Zeithomogenität . . . . .	411
8.5 Tests . . . . .	412
8.6 Spezielle Probleme und Lösungen bei der Anwendung von Markoffketten . . . . .	412
8.7 Einführung unabhängiger Variablen . . . . .	415
8.7.1 Subgruppenmodelle . . . . .	415
8.7.2 Übergangswahrscheinlichkeiten als Funktionen von unabhängigen Variablen . . . . .	415
8.7.3 Interaktive Markoffketten . . . . .	416
8.8 Einführung latenter Klassen . . . . .	416
8.8.1 Mover-Stayer-Modell . . . . .	416
8.8.2 Generelles Modell latenter Zustände . . . . .	417
8.9 Weitere Modelle: zeitkontinuierliche Markoffprozesse . . . . .	417
9. Multivariate „Zeitreihen“- und Panelanalyse mit zeitkontinuierlichen Modellen . . . . .	419
9.1 „Zeitreihenanalyse“ ( $N = 1, T \geq M, M > 1$ ) . . . . .	419
9.1.1 Stochastische Systeme . . . . .	430
9.1.2 Diskrete Approximation des stochastischen zeitkontinuierlichen Modells . . . . .	433
9.1.3 Identifikation und Schätzung des zeitkontinuierlichen Systems . . . . .	436
9.2 Panelanalyse (repeated-measurements) ( $N > M, T \geq 2, M > 1$ ) . . . . .	443
9.2.1 Zeitkontinuierliches Modell . . . . .	443
9.2.2 Diskrete Approximation des stochastischen zeitkontinuierlichen Panelmodells mit LISREL . . . . .	449
9.2.3 Identifikation und Schätzung der zeitkontinuierlichen Panelmodelle . . . . .	450
10. Schlußbemerkungen . . . . .	452

## 4. Kapitel: Statistische Entscheidungstheorie und Bayes-Statistik. Von Dirk Wendt

1. Einleitung: Problemstellung . . . . .	471
1.1 Exkurs über Meßtheorie und Skalierung . . . . .	472
1.2 Schema des Erkenntnisgewinns in einer empirischen Wissenschaft . . . . .	474
2. Klassische Statistik . . . . .	476
2.1 Vorgehensweise der klassischen Statistik . . . . .	476

2.2 *Eigenschaften klassischer Tests* . . . . . 484

2.3 *Zur Frage der Stichprobengröße* . . . . . 486

2.4 *Zur Effektstärke* . . . . . 490

2.5 *Zusammenfassung des klassischen Signifikanztests* . . . . . 491

3. *Sequentielle Testverfahren* . . . . . 494

4. *Likelihood-Quotienten-Test* . . . . . 499

5. *Bayes-Statistik* . . . . . 500

5.1 *Vorgehensweise der Bayes-Statistik* . . . . . 500

5.2 *Robustheit der Schätzung (principle of stable estimation)* . . . . . 501

5.3 *Vergleich mit der klassischen Statistik* . . . . . 503

5.4 *Integration von Daten aus verschiedenen Quellen* . . . . . 504

6. *Parameter-Schätzung* . . . . . 505

6.1 *Lösung der kleinsten Quadrate* . . . . . 505

6.2 *Maximum-Likelihood-Schätzung* . . . . . 506

6.3 *Konjugierte Verteilungen* . . . . . 507

6.4 *Das Principle of Stable Estimation bei der Parameterschätzung* . . . . . 507

7. *Die Erhebung von a-priori-Wahrscheinlichkeiten.* . . . . . 507

8. *Die Bewertung der Ausgänge von Entscheidungen* . . . . . 510

8.1 *Bewertung multiattributiver Ausgänge* . . . . . 515

9. *Entscheidungskriterien* . . . . . 519

10. *Schlußbemerkung* . . . . . 523

5. Kapitel: Computer-Simulation. Von Hans Ueckert

1. *Einleitung.* . . . . . 530

2. *Das Paradigma der Computer-Simulation in der Psychologie* . . . . . 533

2.1 *Zur Klassifikation von Simulationsmodellen* . . . . . 533

2.2 *Programmbeispiel: „Simple Concept Attainment“* . . . . . 536

2.2.1 *Flußdiagrammdarstellung* . . . . . 536

2.2.2 *Das Hauptprogramm (Versuchsablaufprogramm)* . . . . . 538

2.2.3 *Zur „Binnenstruktur“ der Informationsverarbeitung* . . . . . 542

2.2.4 *Die Modellvarianten* . . . . . 544

2.2.5 *Abschließende Funktionsdefinitionen* . . . . . 547

2.3 <i>Diskussion des Programmbeispiels</i> . . . . .	549
2.3.1 Modellcharakteristika . . . . .	549
2.3.2 Nicht-numerisches Programmieren . . . . .	550
2.3.3 „Listenverarbeitung“ . . . . .	551
2.3.4 Modulares Programmieren . . . . .	554
3. Simulationsmodelle und psychologische Theorienbildung . . . . .	555
3.1 <i>Empirische Grundlagenpsychologischer Simulationsmodelle</i> . . . . .	556
3.1.1 Methoden der Datengewinnung . . . . .	556
3.1.2 Möglichkeiten der Datenauswertung . . . . .	558
3.2 <i>Informationelle Produktionssysteme</i> . . . . .	563
3.2.1 Die Modellarchitektur von Produktionssystemen . . . . .	564
3.2.2 Beispiel eines Produktionssystems als Simulationsmodell . . . . .	565
3.2.3 Transparenz und Abbildtreue von Produktionssystemen . . . . .	573
3.3 <i>Das Interpreterproblem von Produktionssystemen</i> . . . . .	575
3.3.1 Lesarten von Produktionsregeln . . . . .	576
3.3.2 Konfliktlösungsstrategien („conflict resolution“) . . . . .	577
3.3.3 Adaptivität (Lernfähigkeit) von Produktionssystemen . . . . .	578
3.3.4 „Bewußtseinsfunktionen“ des Interpreters . . . . .	579
3.4 <i>„Künstliche Intelligenz“ oder: Wie man dem Rechner das Rechnen beibringen kann</i> . . . . .	580
4. Validierung und Anwendbarkeit von Simulationsmodellen . . . . .	587
4.1 <i>Wirklichkeitsbezug und Modellrelationen</i> . . . . .	588
4.1.1 Modellbildung als homomorphe Abbildung . . . . .	588
4.1.2 Grundprobleme der Modellrelationen . . . . .	589
4.1.3 Kommutatives Diagramm . . . . .	591
4.2 <i>Das Eindeutigkeitstheorem von Anderson</i> . . . . .	593
4.3 <i>Empirische Tests von Simulationsmodellen</i> . . . . .	595
4.3.1 Turing-Test . . . . .	596
4.3.2 Protokoll-Trace-Vergleich . . . . .	597
4.4 <i>Nicht-Falsifizierbarkeit von KI-Systemen</i> . . . . .	598
4.4.1 Der strukturalistische Theoriebegriff . . . . .	599
4.4.2 Die logische Komponente der Theorie der Informationsverarbeitung . . . . .	601
4.4.3 Die empirische Komponente der Theorie der Informationsverarbeitung . . . . .	607
4.4.4 Der instrumentelle Gebrauch der Theorie der Informationsverarbeitung . . . . .	609
5. Kommentiertes Literaturverzeichnis . . . . .	610

Autoren-Register . . . . .	617
Sach-Register. . . . .	630

## 1. Kapitel

# Übersicht

*Jürgen Bredenkamp*

### *1. Zur Prüfung sog. Kausalhypothesen*

Selten begnügt man sich in der Psychologie, wie in anderen empirischen Wissenschaftszweigen auch, damit, die korrelativen Beziehungen zwischen verschiedenen Variablen zu konstatieren, ohne zu prüfen, welche Variablen auf andere einwirken. Man intendiert Aussagen wie z.B. „X wirkt sich auf Y aus“. Derartige Interpretationen sind bekanntlich nicht einfach aus den korrelativen Beziehungen zu erschließen. Entweder stellt man aufgrund experimenteller Planungen sicher, daß die Variation von Y als Folge der Variation von X interpretiert werden kann, oder es wird, falls experimentelles Handeln nicht möglich ist, ein Modell über die Beeinflussungsrichtung zwischen den Variablen aufgestellt, das auch dann prüfbar ist, wenn es sich um reine Korrelationsforschung handelt.

Dem Experiment, dessen Planung und Auswertung ausführlich durch Hager und Westermann in diesem Band behandelt werden, kommt unter den verschiedenen Methoden zur Überprüfung sog. Kausalhypothesen - dieser Begriff wird noch präzisiert werden - eine Sonderstellung zu. Der Experimentator selbst stellt verschiedene Bedingungen her (er „manipuliert“ eine Variable) und beobachtet die Auswirkungen dieser „unabhängigen“ auf eine andere „abhängige“ Variable. Dadurch wird eine zeitliche Abfolge unabhängige Variable „X“  $\rightarrow$  abhängige Variable „Y“ hergestellt, und die korrelative Beziehung zwischen beiden Größen kann nicht derart interpretiert werden, daß sich Y auf X auswirkt. Allerdings reicht das bisher geschilderte Vorgehen noch nicht aus, um behaupten zu können, daß X auf Y einwirkt. Mit X könnten eine oder mehrere Störvariablen „St“ korreliert sein, die „in Wirklichkeit“ für den beobachteten Zusammenhang zwischen X und Y verantwortlich sind. Um die Möglichkeit einer derartigen Scheinbeziehung zwischen X und Y, wie sie in Abb. 1 dargestellt ist, zu reduzieren, bedient man sich in der experimentellen Psychologie verschiedener Kontrollverfahren; in der Terminologie Campbells und Stanleys (1963) kommt einem Experiment um so größere interne



Validität zu, je geringer die Möglichkeit zu derartigen Scheinbeziehungen ist. Wichtig ist vor allem die zufällige Zuweisung der Probanden auf die experimentellen Bedingungen (Randomisierung), die unerlässlich ist, um den Erwartungswert der Korrelation zwischen Personmerkmalen und X Null werden zu lassen. Neben der Randomisierung kommen vor allem die Konstanthaltung und die von X unabhängige systematische Variation bekannter Störgrößen in Frage. Durch diese Variation wird eine zweite „unabhängige“ Variable zur Kontrolle eingeführt, und es ist jetzt prüfbar, ob X auf allen Stufen von St denselben Einfluß auf Y ausübt (die statistische Interaktion zwischen X und St ist Null) oder ob dieser Einfluß von St abhängig ist (X und St interagieren). In der experimentellen Psychologie wird, wie gesagt, meistens so verfahren, daß X und St nicht korrelieren. Unterstellt man seinen Daten das lineare Modell einer multiplen Regression (vgl. dazu Schubö et al. in Band 4 dieser Enzyklopädie), so läßt sich das zugrundeliegende Beeinflussungsmodell wie in Abb. 2 darstellen, wobei vorausgesetzt ist, daß X, St und die Interaktionsvariable XSt wechselseitig nicht miteinander korrelieren (diese Variablen sind deshalb in Abb. 2 nicht durch Pfeile verbunden).

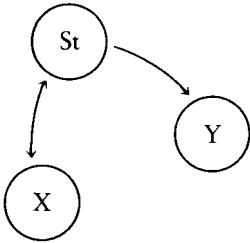


Abb. 1: Der gekrümmte Doppelpfeil weist auf eine korrelative Beziehung zwischen X und St hin, die nicht daraufhin analysiert werden kann, welche Variable unabhängig in bezug auf die andere ist. St beeinflusst Y direkt, während der Zusammenhang zwischen X und Y zum Schein besteht: X und Y korrelieren nur deshalb, weil X mit einer Variablen korreliert, die einen direkten Einfluß auf Y ausübt.

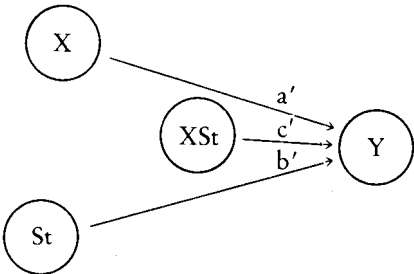


Abb. 2:

In diesem Graphen symbolisieren  $a'$ ,  $b'$ ,  $c'$  die Einflußgrößen der Variablen  $X$ ,  $St$  und  $XSt$  auf  $Y$ . Meistens wird mittels der Varianzanalyse geprüft, ob  $a'$ ,  $b'$  und  $c'$  Null sind (die Varianzanalyse ist ein Spezialfall der multiplen Regressionsanalyse, und  $a'$ ,  $b'$ ,  $c'$  sind die (multiplen) Korrelationen zwischen  $X$  und  $Y$ ,  $St$  und  $Y$  sowie  $XSt$  und  $Y$ ). Ist  $c' \neq 0$  und  $a' = 0$ , spricht man von einer für  $X$  disordinalen Interaktion (vgl. Bredenkamp, 1980): Die Richtung der Beeinflussung von  $X$  auf  $Y$  hängt von dem Wert der Störvariablen ab. Wenn  $c' \neq 0$  und  $a' \neq 0$ , ist für die experimentelle Variable  $X$  aus diesen Informationen nicht ableitbar, ob  $X$  mit  $St$  derart interagiert, daß die gleiche Beeinflussungsrichtung von  $X$  auf  $Y$  unter allen Werten von  $St$  vorliegt (Interaktion für  $X$  ist ordinal). Sollte die Interaktion für  $X$  disordinal sein, kann nicht von einer Kausalbeziehung zwischen  $X$  und  $Y$  gesprochen werden. Diese kann nur dann vorliegen, wenn  $a' \neq 0$  und  $c' = 0$ , oder wenn  $a' \neq 0$  und  $c' \neq 0$ , zusätzlich aber gezeigt worden ist, daß die Interaktion an der Beeinflussungsrichtung von  $X$  auf  $Y$  nichts ändert (vgl. dazu Bredenkamp, 1982).

Mit der Nennung dreier Techniken ist das Reservoir notwendiger Kontrollen, damit das Experiment dem Anspruch als Prüféxperiment von Kausalhypothesen gerecht werden kann, nicht ausgeschöpft. Ausführlich hierüber informieren Hager und Westermann in diesem Band (vgl. auch Bredenkamp, 1980). Später werden wir noch auf einen Aspekt der Kontrolle zu sprechen kommen, welche die Validität des statistischen Schlusses sichern soll.

In Abb. 1 und Abb. 2 sind sog. rekursive Systeme dargestellt, die dadurch ausgezeichnet sind, daß in den Graphen keine Zyklen auftreten. Es gibt keinen Pfad von einer Variablen zu einer anderen Variablen, von der aus man wieder zum Ausgangspunkt zurückkommt. Der Doppelpfeil in Abb. 1 besagt nur, daß die Richtung der Beziehung zwischen  $X$  und  $St$  nicht analysiert wird. In einem nicht-rekursiven System würde dagegen in Abb. 1 ein Pfeil von  $X$  nach  $St$  und ein anderer Pfeil von  $St$  nach  $X$  laufen. Derartige Systeme werden hier nicht betrachtet (vgl. dazu Hummell und Ziegler, 1976). In rekursiven Systemen heißen solche Variablen, von denen nur Pfeile ausgehen, exogen, während Variablen, bei denen wenigstens ein Pfeil ankommt, endogen genannt werden. Üblicherweise sind also die unabhängigen Variablen eines Experiments exogene, die abhängigen Variablen endogene Variablen innerhalb eines rekursiven Systems; allerdings finden sich, wie noch gezeigt wird, auch Beispiele, in denen manche unabhängige Variable endogen ist. Ein rekursives System soll nur dann kausal heißen, wenn es entweder keine Interaktionsvariablen enthält, oder wenn gezeigt werden kann, daß die Interaktionsvariablen die Beeinflussungsrichtung der interessierenden Variablen auf andere nicht modifizieren.

Die Formulierung rekursiver Systeme ist selbstverständlich nicht auf die experimentelle Psychologie beschränkt, sondern auch in der Korrelationsforschung möglich, wobei unter Korrelationsforschung die Analyse korrelativer Bezie-

hungen zwischen Variablen aus einer Untersuchung zu verstehen ist, der nicht die Merkmale „Manipulation“ wenigstens einer Variablen und „Randomisierung“ zukommen (s.o.). Z.B. könnten für die Variablen „Allgemeine Intelligenz“ (AI), „Intelligenz in der Wahrnehmung des Lehrers“ (IW) und „Zeugnisnote“ (ZN) die rekursiven Systeme in Abb. 3a und Abb. 3b formuliert werden (vereinfachtes Beispiel aus Brandstädter und Bernitzke, 1976).

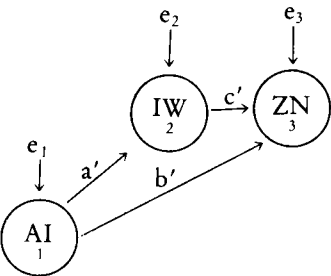


Abb. 3a:

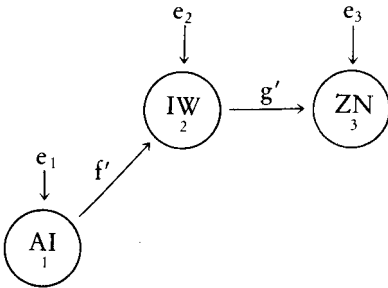


Abb. 3b:

In Abb. 3a wird ein direkter und ein über IW vermittelter Einfluß von AI auf ZN postuliert, während in Abb. 3b nur ein vermittelter Einfluß angenommen wird. Die impliziten Variablen e bezeichnen durch die Problemstellung nicht berücksichtigte Größen, durch deren Einführung man formal eine Schließung des Systems erreicht. Systeme wie in Abb. 3a und Abb. 3b heißen lineare Kausalstrukturen, wenn die Beziehungen zwischen den Variablen ausschließlich linear sind; nur derartige Systeme sollen hier betrachtet werden.

Das System in Abb. 3a heißt vollständig, da Wirkungen zwischen allen Variablenpaaren postuliert werden. Dagegen heißt das System in Abb. 3b unvollständig, da es nicht alle  $\binom{m}{2}$  Beziehungen zwischen den m expliziten Variablen enthält.

Von einer kausalen Wirkung von X auf Y innerhalb einer gegebenen Kausalstruktur soll nur dann gesprochen werden, wenn von X auf Y nachweisbar direkte und/oder indirekte Einflüsse bestehen (und wenn die aufgrund der Kausalstruktur vorhergesagten Korrelationen zwischen den Variablen den empirisch ermittelten Korrelationen entsprechen; vgl. Hummell und Ziegler, 1976). Damit sind zwei Prüfbedingungen angesprochen, auf die jetzt eingegangen werden soll. Zuvor sei betont, daß die Aussage „X wirkt kausal auf Y“ immer nur innerhalb eines bestimmten rekursiven Systems gilt oder nicht zutrifft.

Das übliche Experiment versucht durch die Anwendung von Kontrolltechniken zu erreichen, daß der gesamte kausale Einfluß von X auf Y als ausschließ-

lich direkter Einfluß analysiert werden kann. Dieses Vorgehen ist in dem Maße erfolgreich, wie gewährleistet ist, daß durch die Manipulation von X nicht andere Variablen, die auf Y einwirken, verändert werden (interne Validität). Die Prüfung der Aussage „X wirkt kausal auf Y“ geschieht derart, daß statistisch die Hypothese  $a'=0$  getestet wird (äquivalent dem Test der Hypothese, daß die Korrelation zwischen X und Y Null ist). Wenn diese Hypothese  $a'=0$  abgelehnt werden kann, gilt die Aussage „X wirkt auf Y kausal“ als vorläufig bewährt, wobei noch durch die Einführung von sog. Maßen der praktischen Signifikanz (vgl. dazu Hager und Westermann in diesem Band) zusätzlich gefordert werden kann, daß a, die Schätzung für  $a'$ , einen gewissen Wert übersteigt. Vorauszusetzen ist außerdem immer, daß, bezogen auf Abb. 2,  $c'=0$  oder die Interaktion zwischen X und St nichts an der Beeinflussungsrichtung von X auf Y unter verschiedenen Werten von St ändert. Wenn jedoch die Hypothese  $a'=0$  angenommen werden muß, liegt eine notwendige, aus verschiedenen Gründen jedoch noch nicht hinreichende Bedingung (vgl. dazu Bredenkamp, 1980) für den Schluß or, X wirke nicht kausal auf Y.) Auf statistische Erfordernisse beim Test derartiger Hypothesen kommen wir in Abschnitt 2 zu sprechen. Weitere Prüfungen sind nicht möglich, sofern im rekursiven System nur direkte Wirkungen postuliert werden und keine Vorinformationen über die Größe der Einflüsse  $a'$ ,  $b'$ ,  $c'$  existieren.

In der Korrelationsforschung lassen sich ebenfalls Kausalhypothesen prüfen. Für die Prüfung von Kausalhypothesen ist nicht die experimentelle Kontrollmöglichkeit entscheidend, sondern die Konzeption eines kausalen rekursiven Systems. Die Kontrolle im Experiment soll nur bewirken, daß allein direkte kausale Einflüsse analysiert werden können. In der Korrelationsforschung dagegen wird wegen eingeschränkter Kontrollmöglichkeiten von vornherein von direkten und indirekten Einflüssen ausgegangen. Die rekursive Pfadanalyse ist das Analyseverfahren, um Kausalhypothesen in der Korrelationsforschung zu überprüfen. Wir gehen hier nur auf die rekursive lineare Pfadanalyse ein (vgl. etwa Brandtstädter und Bernitzke, 1976; Hummell und Ziegler, 1976; Kerlinger und Pedhazur, 1973). Die Anwendung dieses Verfahrens setzt voraus:

- (1) Die Konzeption eines linearen rekursiven Systems.
- (2) Die expliziten Variablen sind metrisch und fehlerfrei gemessen (zur Lockerung dieser Annahmen siehe Hummell und Ziegler, 1976).
- (3) die impliziten Variablen e (vgl. Abb. 3), welche die endogenen Variablen beeinflussen, sind mit keiner anderen Variablen des Systems korreliert.
- (4) Alle Beziehungen zwischen den Variablen sind ausschließlich linear und additiv.
- (5) Interaktionseffekte gibt es nicht.

---

<sup>1</sup>) Diese Darstellungen unterstellen eine multiple Regressionsanalyse experimenteller Daten, die immer dann möglich ist, wenn varianzanalytische Strukturmodelle mit festen Effekten den Daten zugrunde liegen.

Können diese Annahmen für die Beispiele in Abb. 3a und 3b als gültig unterstellt werden, und liegen die Variablen dieser Systeme als Standardwerte  $Z$  mit dem Mittelwert 0 und der Standardabweichung 1 vor, läßt sich für die Daten des Modells in Abb. 3a schreiben:

- 1)  $Z_1 = e_1$
- 2)  $Z_2 = aZ_1 + e_2$
- 3)  $Z_3 = bZ_1 + cZ_2 + e_3$ ,

wobei  $a$ ,  $b$ ,  $c$  Schätzungen der Parameter  $a'$ ,  $b'$ ,  $c'$  sind. Multipliziert man Gleichung 2) mit  $Z_1$ , summiert über alle  $N$  Probanden und dividiert durch  $N$ , erhält man:

$$r_{12} = a.$$

Die gleiche Verfahrensweise führt zu:

$$\begin{aligned} r_{13} &= b + cr_{12} \\ r_{23} &= br_{12} + c, \end{aligned}$$

wobei die beiden letzten Gleichungen zur Schätzung der Pfadkoeffizienten  $b$  und  $c$  verwendet werden:

$$b = \frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2} \quad c = \frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}.$$

Das in Abb. 3a veranschaulichte System ist vollständig, und wie gezeigt worden ist, werden alle möglichen Korrelationen benötigt, um die Einflußgrößen (Pfadkoeffizienten) zu schätzen. Sofern also keine Vorinformationen über diese Größen vorliegen, lassen sich keine Korrelationen zwischen den Variablen eines vollständigen Systems prognostizieren. Dies trifft auch für das Experiment zu, soweit es bisher besprochen wurde. Im Unterschied zum Experiment wird aber in einer vollständigen linearen Kausalstruktur die kausale Wirkung auf direkte und indirekte Einflüsse zurückgeführt. Dies wird sichtbar, wenn in den Gleichungen für  $r_{13}$  und  $r_{23}$   $a$  für  $r_{12}$  eingesetzt wird:

$$\begin{aligned} r_{13} &= b + ac \\ r_{23} &= c + ab. \end{aligned}$$

$r_{13}$  ist also eine Schätzung des kausalen Einflusses von  $Z_1$  auf  $Z_3$  innerhalb der linearen Kausalstruktur in Abb. 3a, der sich aus einer direkten Wirkung  $b$  und einer indirekten Wirkung  $ac$  zusammensetzt. Auch  $Z_2$  hat einen direkten ( $c$ ) Einfluß auf  $Z_3$ , aber keinen indirekten, da von  $Z_2$  kein weiterer Pfad nach  $Z_3$  führt:  $ab$  gibt hier die in  $r_{23}$  enthaltene Scheinbeziehung an.

Vollständige Modelle nun lassen sich, genau wie dem Experiment zugrunde liegende Modelle, nur derart prüfen, daß für die Pfadkoeffizienten getestet

wird, ob sie statistisch gesehen von Null abweichen. Ergibt sich, daß alle Pfadkoeffizienten von Null abweichen, ist dies noch keine Bestätigung für das geprüfte vollständige Modell. Jede andere vollständige lineare Kausalstruktur mit den gleichen expliziten Variablen würde gleichermaßen gut für die Daten passen. Wenn jedoch zusätzlich eine zeitliche Ordnung zwischen den Variablen postuliert werden kann, derart z.B., daß wie in Abb. 3a AI vor IW und IW vor ZN steht, könnte das Modell als bewährt angesehen werden. Ist jedoch wenigstens einer der Pfadkoeffizienten Null, kann ein unvollständiges Modell formuliert werden, in dem nur einige Korrelationen zur Schätzung der Pfadkoeffizienten benötigt werden. In diesem Fall lassen sich andere Korrelationen prognostizieren, und die Prognosen können mit den tatsächlichen erhaltenen Daten verglichen werden.

Wenn die empirisch erhaltenen Korrelationen zu Abb. 3a bei  $N = 100$  Probanden  $r_{12} = 0.40$ ,  $r_{13} = 0.25$  und  $r_{23} = 0.60$  betragen, läßt sich ermitteln:

$$a = 0.40 \qquad b = 1/84 \qquad c = 50/84.$$

Da  $b$  nahezu Null ist, wurde das Modell aus Abb. 3b formuliert, dessen Strukturgleichungen für Stichprobendaten lauten:

$$\begin{aligned} 4) \quad Z_1 &= e_1 \\ 5) \quad Z_2 &= fZ_1 + e_2 \\ 6) \quad Z_3 &= gZ_2 + e_3. \end{aligned}$$

Durch Multiplikation von Gleichung 5) mit  $Z_1$ , Aufsummierung und Division durch  $N$  ergibt sich:

$$r_{12} = f.$$

Ein entsprechendes Vorgehen führt zu:

$$\begin{aligned} r_{13} &= gr_{12} = fg. \\ r_{23} &= g. \end{aligned}$$

$g$  kann also auf zweierlei Weise geschätzt werden. Aus  $g = r_{23}$  ergibt sich die Prognose für eine Korrelation, die zur Schätzung der Pfadkoeffizienten nicht benötigt wird:

$$r_{13}^* = r_{12}r_{23} = 0.24.$$

Verglichen mit dem tatsächlichen Wert  $r_{13} = 0.25$  ist die Prognose recht genau, und man kann auf einen Signifikanztest verzichten. Ist die Übereinstimmung nicht derart deutlich, muß aus  $r_{13}^* = r_{12}r_{23}$  eine Prognose abgeleitet werden, die sich statistisch prüfen läßt. Würde man folgern, daß die Partialkorrelation

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2} \sqrt{1-r_{23}^2}}$$

Null sein muß, so ginge in diese Folgerung die modellunverträgliche Annahme ein, daß  $Z_2$  auf  $Z_1$  und  $Z_3$  einwirkt (vgl. Abb. 3b). Folgert man jedoch, daß die semipartielle Korrelation

$$r_{1(3.2)} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{23}^2}}$$

Null ist, so liegt eine modellverträgliche Prognose vor, in die die Annahme eingeht, daß  $Z_2$  sich auf  $Z_3$  auswirkt. Diese Prognose läßt sich statistisch prüfen.

Nunmehr soll die rekursive Pfadanalyse auf ein komplexeres Beispiel angewendet werden. Dieses Beispiel ist der experimentellen Psychologie entnommen (ein komplexes Beispiel aus der Korrelationsforschung findet sich bei Brandtstädter, 1976). Auch im Experiment ist es nicht immer möglich, die verschiedenen Variablen eines rekursiven Systems unabhängig voneinander so zu variieren, daß nur direkte kausale Einflüsse analysiert werden können. Will man z. B. den Einfluß der Bildhaftigkeit des Lernmaterials auf die Gedächtnisleistung prüfen, so sind andere Variablen mit der Bildhaftigkeit konfundiert. So entsteht etwa beim Lernen von Sätzen das Problem, ob die Bildhaftigkeit oder die Verständlichkeit der Sätze die Gedächtnisleistung determiniert. Wipich und Bredenkamp (1979) haben argumentiert, daß die Bildhaftigkeit der Sätze die Variable „Verständlichkeit“ beeinflusst und nicht umgekehrt. Nimmt man die Annahmen hinzu, daß die Bildhaftigkeit der Substantive in Subjekt- und Objektposition die Bildhaftigkeit des Satzes determinieren, so läßt sich das Modell in Abb. 4 formulieren:

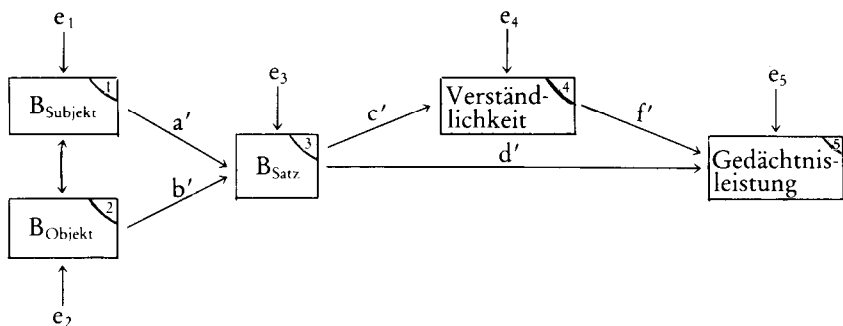


Abb. 4:

Die Gedächtnisleistung wurde unter fünf verschiedenen experimentellen Bedingungen erhoben, die erst später beschrieben werden sollen. Entsprechend wird diese Variable in Tab. 1, die die empirisch erhaltenen Korrelationen wiedergibt, mit 5a bis 5e bezeichnet. Folgende Strukturgleichungen lassen sich nach dem Modell in Abb. 4 für die Daten schreiben:

- 7)  $Z_1 = e_1$
- 8)  $Z_2 = e_2$
- 9)  $Z_3 = aZ_1 + bZ_2 + e_3$
- 10)  $Z_4 = cZ_3 + e_4$
- 11)  $Z_5 = dZ_3 + fZ_4 + e_5$

Tabelle 1:

	1	2	3	4
1				
2	0.93			
3	0.96	0.96		
4	0.90	0.90	0.93	
5a	0.78	0.74	0.73	0.75
5b	0.82	0.78	0.80	0.74
5c	0.91	0.92	0.93	0.87
5d	0.21	0.21	0.22	0.26
5e	0.86	0.87	0.88	0.83

Durch Multiplikation der Gleichung 9) mit  $Z_1$  (bzw.  $Z_2$ ), Aufsummierung und Division durch N erhält man:

$$\begin{aligned} r_{13} &= a + br_{12} \\ r_{23} &= ar_{12} + b \end{aligned}$$

Diese beiden Gleichungen genügen, um  $a = 0.50$  und  $b = 0.50$  zu schätzen.<sup>2)</sup> Weiterhin läßt sich zeigen, daß

$$r_{34} = c = 0.93.$$

Da  $r_{14}$  und  $r_{24}$  für keine Schätzung benötigt werden, lassen sie sich prognostizieren:

$$\begin{aligned} r_{14} &= cr_{13} = ac + bcr_{12} = 0.89 \text{ (tatsächlicher Wert: } 0.90) \\ r_{24} &= cr_{23} = acr_{12} + bc = 0.89 \text{ (tatsächlicher Wert: } 0.90). \end{aligned}$$



Weiterhin läßt sich zeigen, daß:

$$\begin{aligned} r_{15} &= dr_{13} + fr_{14} = ad + bdr_{12} + acf + bcfr_{12} \\ r_{25} &= dr_{23} + fr_{24} = adr_{12} + bd + acfr_{12} + bcf \\ r_{35} &= d + fr_{34} = d + cf \\ r_{45} &= dr_{34} + f = cd + f. \end{aligned}$$

Für zwei unbekannte Größen d und f stehen vier Gleichungen zur Verfügung; zwei werden jedoch nur benötigt. Entscheidet man sich dafür, die Gedächtnisleistung aus der Bildhaftigkeit des Subjekts und Objekts ( $Z_1, Z_2$ ) prognostizieren zu wollen, schätzt man aus den Gleichungen für  $r_{35}$  und  $r_{45}$  die Größen d und f. Tab. 2 enthält die Werte für d, f,  $r_{15}^*$  und  $r_{25}^*$ . Beim Vergleich mit den tatsächlich erhaltenen Korrelationen aus Tab. 1 fällt die gute Übereinstimmung auf. Ferner fällt auf, daß vermutlich nur für Variable 5a das vollständige Modell aus Abb. 4 benötigt wird, während für die Variablen 5b, 5c und 5e das Modell in Abb. 5a und für Variable 5d das Modell in Abb. 5b ausreicht.

Tabelle 2:

	d	f	$r_{15}$	$r_{25}$
Variable 5a	0.24	0.53	0.70	0.70
5b	0.83	-0.03	0.77	0.77
5c	0.89	0.04	0.89	0.89
5d	-0.16	0.41	0.21	0.21
5e	0.80	0.09	0.85	0.85

Aus den Strukturgleichungen für Modell 5a lassen sich dieselben Pfadkoeffizienten a, b, c wie für das Modell 4 gewinnen, außerdem ist jetzt  $d = r_{35}$ . Nunmehr läßt sich prognostizieren für Variable 5:

$$\begin{aligned} r_{15}^* &= dr_{13} = ad + bdr_{12} \\ r_{25}^* &= dr_{23} = adr_{12} + bd \\ r_{45}^* &= dr_{34} = cd. \end{aligned}$$

\*) Bei der Schätzung der Pfadkoeffizienten können sich erhebliche Fehler einstellen, wenn die Variablen so hoch miteinander korrelieren, wie das in diesem Beispiel der Fall ist. Um derartige Schätzfehler zu vermeiden, sollte man deshalb mit einem Modell arbeiten, das nur die Variablen  $Z_1, Z_2$ , und  $Z_5$  erhält. Nur zur Demonstration eines komplexeren Modelies wird hier auf diese empfehlenswerte Reduktion verzichtet.

Grundsätzlich ist zur Prognose der Korrelationen folgendes zu sagen:  $r_{ij}$ , mit  $i < j$  wird immer gewonnen, indem die Gleichung für  $Z_i$  mit  $Z_j$  multipliziert wird. Die Prognose erfolgt durch Verknüpfung der Pfadkoeffizienten und nicht analysierter Korrelationen (in diesem Beispiel  $r_{12}$ ).

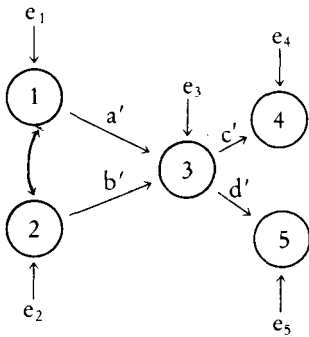


Abb. 5a:

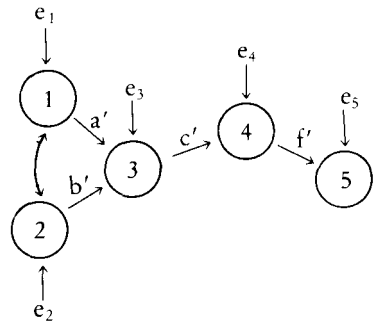


Abb. 5b:

Dieses Modell, das eine Beziehung weniger als Modell 4 postuliert, ermöglicht auch eine Prognose mehr. Es ergeben sich, wie ein Vergleich der Tab. 3 mit Tab. 1 zeigt, gute Übereinstimmungen.

Tabelle 3:

		$r_{15}$	$r_{25}$	$r_{35}$
Variable	5b	0.77	0.77	0.74
	5c	0.89	0.89	0.86
	5e	0.84	0.84	0.82

Aus den Strukturgleichungen für Modell 5b lassen sich wiederum dieselben Pfadkoeffizienten  $a$ ,  $b$ ,  $c$  schätzen, und  $f$  ergibt sich zu  $r_{45}$ . Prognostizieren lassen sich für Variable 5:

$$\begin{aligned} r_{15}^* &= fr_{14} = fcr_{13} = acf + bcfr_{12} \\ r_{25}^* &= fr_{24} = fcr_{23} = acfr_{12} + bcf \\ r_{35}^* &= fr_{34} = fc. \end{aligned}$$

Die Prognosen und tatsächlichen Werte für Variable 5d enthält Tab. 4.

Tabelle 4:

$r_{15}^* = 0.23$	$r_{25}^* = 0.23$	$r_{35}^* = 0.24$
$r_{15} = 0.21$	$r_{25} = 0.21$	$r_{35} = 0.22$

Kommen wir nun auf die Interpretation zu sprechen. Als zentral soll die Aufklärung des Zusammenhangs zwischen der Bildhaftigkeit der Sätze und der

Gedächtnisleistung angesehen werden. Das Modell in Abb. 4 postuliert einen direkten Einfluß, dessen Größe durch den Pfadkoeffizienten  $d$  geschätzt wird, sowie einen indirekten Einfluß cf. Dieses Modell wurde für Variable 5a akzeptiert, wobei  $d = 0.24$  und  $cf = 0.49$ . Bei Variable 5a handelt es sich um die Reproduktionsleistung nach intentionalem Lernen: Die Vpn wußten, daß die Sätze reproduziert werden mußten. In dieser Situation besteht also ein direkter und ein über die Verständlichkeit vermittelter Einfluß der Bildhaftigkeit der Sätze auf die Reproduktionsleistung. Für die Variablen 5b, 5c und 5e wurde das Modell in Abb. 5a konzipiert, nachdem sich bei der Berechnung der Pfadkoeffizienten für das komplexere Modell in Abb. 4 herausgestellt hatte, daß  $f$  nahe Null liegt; auch die Korrelationen ließen sich mit Hilfe dieses Modells gut reproduzieren. Wenn man will, kann man mit Hilfe statistischer Hypothesentests den Vergleich beider Modelle vornehmen. Da im Modell der Abb. 4 die Variable „Gedächtnisleistung“ durch die Prädiktoren  $Z_3$  und  $Z_4$ , im Modell der Abb. 5a jedoch nur durch den Prädiktor  $Z_3$  determiniert wird, sollte statistisch gesehen  $Z_4$  keine zusätzliche Varianzaufklärung leisten:  $R_{5,3,4}^2 - r_{53}^2$  sollte nicht signifikant von Null abweichen. Dieser Betrag gibt die Größe der quadrierten semipartiellen Korrelation zwischen Gedächtnisleistung und Verständlichkeit wieder, nachdem aus dieser Variablen der Einfluß der Bildhaftigkeit der Sätze ausgeschaltet wurde. Für die Variablen 5b, 5c, 5e ist dieser Betrag mit 0.01, 0.01 und 0.01 sehr klein und nicht signifikant. Die Tests, die  $R_{5,3,4}^2 - r_{53}^2$  auf Signifikanz prüfen, sind Prüfungen der Bedeutsamkeit des Pfadkoeffizienten  $f'$  aus Modell 4. Das Modell in Abb. 5a postuliert nur einen direkten Einfluß der Bildhaftigkeit der Sätze auf die Gedächtnisleistung. Bei den Variablen 5b und 5c handelt es sich ebenfalls um Gedächtnisleistungen nach dem intentionalen Lernen. Allerdings hatten die Pbn Zusatzaufgaben während des Lernens zu verrichten, die nach dem levels of processing-Ansatz verschiedene Verarbeitungsformen induzieren sollten: Für Variable 5b sollte die Bildhaftigkeit eine geringere Rolle als für Variable 5c spielen, was auch, wenn man die beiden Korrelationen  $r_{35}$  vergleicht, der Fall war. Überraschend ist allerdings der Befund, daß auch für Variable 5b nur ein direkter Einfluß der Bildhaftigkeit besteht. Verglichen mit Variable 5a führt also beim intentionalen Lernen jede der verwendeten Zusatzaufgaben zu einem direkten Einfluß der Bildhaftigkeit; nur ohne Zusatzaufgabe besteht auch ein indirekter Einfluß. Nach inzidentellem Lernen (Variablen 5d und 5e) besteht nur nach „kognitiv tiefer“ Verarbeitung (Variable 5e) ein direkter Effekt der Bildhaftigkeit, während nach „oberflächlicher“ Verarbeitung (Variable 5d) allein ein indirekter Effekt besteht. Die Angemessenheit des Modells in Abb. 5b für Variable 5d kann dadurch getestet werden, daß wegen  $r_{35}^* = cf = r_{34} \cdot r_{45}$  folgt: Die semipartielle Korrelation  $r_{3(5,4)}$  weicht nicht signifikant von Null ab. Der Wert für diese Korrelation beträgt für Variable 5d -0.02 und ist insignifikant.

Die vorgetragenen Ergebnisse haben u.E. bisher nicht gesehene Konsequenzen beim Versuch der Zusammenführung des levels of processing-Ansatzes

mit der „imagery“-Theorie Paivios, auf die hier nicht einzugehen ist. Einige dieser Ergebnisse kamen unerwartet; ihre Interpretation ist im Rahmen der behandelten Modelle gültig. Die Richtigkeit derartiger Modelle läßt sich selbstverständlich nicht beweisen. Es sind immer andere Modelle denkbar, die ebenfalls gut für die Daten passen. Die Pfadanalyse ist, wie Kerlinger und Pedhazur (1973) zu Recht feststellen, ein Verfahren, das die Zurückweisung unhaltbarer Kausalmodelle eher als die Bestätigung eines von verschiedenen rivalisierenden Kausalmodellen leistet. Die Pfadanalyse ist auch kein Verfahren, das zu Modellen führt! Die Theorie- oder Modellbildung wird dem Forscher nicht abgenommen. Sie ist unabdingbare Voraussetzung für die Anwendung der Pfadanalyse, die ein Verfahren zur Prüfung von Theorien ist und möglicherweise Hinweise, wie gezeigt wurde, dafür gibt, an welchen Stellen eine Theorie zu modifizieren ist. Leider sind die Möglichkeiten der Pfadanalyse zur Prüfung kausaler Hypothesen in der Korrelationsforschung bisher viel zu selten genutzt worden. Einen methodisch völlig anders gearteten Ansatz zur Analyse von Kausalbeziehungen beschreibt Lehmann (1980).

Das behandelte Beispiel aus der experimentellen Psychologie zeigt Analysemöglichkeiten auf, wenn unabhängige Variablen konfundiert sind: Wenigstens eine dieser Variablen wird als endogen konzipiert. Dieser Fall liegt bei der Anwendung der Kovarianzanalyse generell vor: Im Rahmen eines rekursiven linearen Modells betrachtet wird angenommen, daß die Kovariate  $X$  sich auf die endogene unabhängige Variable  $U$  und die abhängige Variable  $Y$  auswirkt (vgl. Abb. 6a). Die Kovarianzanalyse prüft, ob die Einflußgröße  $c'$  Null ist, indem die semipartielle Korrelation zwischen  $Y$  und  $U$ , aus der der Einfluß der Kovariaten eliminiert wird, auf Signifikanz getestet wird. Dieses Vorgehen ist sinnvoll, da  $U$  nur einen direkten kausalen Einfluß auf  $Y$  ausübt. In der Korrelation zwischen  $U$  und  $Y$  ist auch noch eine Scheinbeziehung zwischen beiden Variablen enthalten ( $a'b'$ ), und deshalb interessiert allein die Größe von  $c$ , die den gesamten kausalen Einfluß von  $U$  auf  $Y$  erfaßt (übrigens läßt man in der experimentellen Psychologie die Scheinbeziehung durch zufällige Zuweisung der Probanden auf die experimentellen Bedingungen von vornherein Null werden, da in diesem Fall  $a'=0$ ).

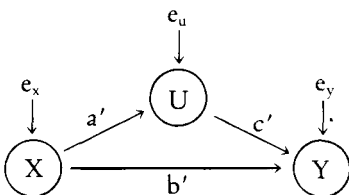


Abb. 6a: Modell für die Kovarianzanalyse.

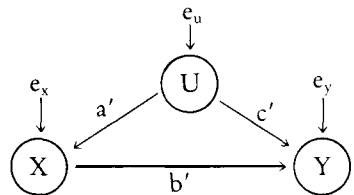


Abb. 6b: Unzutreffendes Modell für die Kovarianzanalyse.

Nun finden sich häufig Beispiele für die Verwendung einer Kovarianzanalyse in der Situation, die in Abb. 6b dargestellt ist. In diesem Fall wirkt sich die unabhängige Variable auf die Kovariate aus, und man möchte wissen, ob ein Einfluß von U auf Y besteht, nachdem die Unterschiede in der Kovariaten statistisch ausgeglichen wurden. Die Anwendung der Kovarianzanalyse ist in diesem Fall nicht gerechtfertigt.  $r_{uy}$  gibt in Abb. 6b den gesamten Einfluß von U auf Y wieder. Will man prüfen, ob ein direkter Einfluß von U auf Y besteht, testet man die semipartielle Korrelation  $r_{U(Y.X)}$  auf Signifikanz.

## 2. Zum Problem der Validität des statistischen Schlusses

Kausalmodelle werden in der experimentellen Psychologie und der Korrelationsforschung statistisch überprüft. Diese Feststellung führt zur Frage der Validität des statistischen Schlusses. In der experimentellen Psychologie implizieren psychologische Hypothesen häufig die zur Nullhypothese eines statistischen Tests alternative Hypothese, so daß auf das Zutreffen der Nullhypothese erkannt werden können muß, damit die psychologische Hypothese falsifizierbar ist. Dies ist nur dann möglich, wenn die Wahrscheinlichkeit des  $\beta$ -Fehlers ebenso wie die des  $\alpha$ -Fehlers kontrolliert wird. Es wird also empfohlen, eine minimale Effektgröße festzulegen, die mit der Wahrscheinlichkeit  $1 - \beta$  entdeckt werden soll, und  $\alpha$  und  $\beta$  als kleine Beträge festzulegen (Bredenkamp, 1969, 1972). Für verschiedene statistische Verfahren läßt sich dann der benötigte Stichprobenumfang bestimmen. Die bei Cohen (1977) publizierten Tabellen erleichtern diese Bestimmung wesentlich. Das hier nur andeutungsweise vorgestellte Verfahren (ausführlich dazu Bredenkamp, 1980; Hager und Westermann in diesem Band) ist auch dann notwendig, wenn die statistische Nullhypothese durch die psychologische Hypothese impliziert wird, um zu viele unverdiente Bestätigungen der psychologischen Hypothesen zu vermeiden. Außerdem ist eine derartige Planung auch für die pfadanalytische Überprüfung rekursiver linearer Systeme vonnöten, wenn auch hier statistische Tests zur Prüfung des Modells durchgeführt werden und man sich nicht mit der augenscheinlichen Übereinstimmung der reproduzierten mit den tatsächlich erhaltenen Korrelationen begnügt. Schwierigkeiten bereitet dieses Verfahren insofern, als zu Beginn eines Forschungsprogramms die Festlegung einer minimalen Effektstärke unmöglich erscheint. Diesem Problem ist man jedoch nicht entronnen, wenn man auf eine derartige Festlegung verzichtet, da man durch die unbegründete Wahl des Stichprobenumfangs auf eine minimale Effektstärke festgelegt worden ist, die mit vorgegebenen Wahrscheinlichkeiten  $\alpha$  und  $1 - \beta$  entdeckt werden kann. In diesem Fall bleibt auch die Frage, ob Versuchsergebnisse durch andere Forscher repliziert worden sind, unentscheidbar (vgl. Bredenkamp, 1980). Wenn jedoch anerkannt wird, daß psychologische Hypothesen oder Kausalmodelle nicht aufgrund einzelner Untersuchungen, sondern ganzer Forschungsprogramme falsifiziert werden kön-

nen, läßt sich das Problem der Festlegung der Effektgröße entschärfen. Legt man diese „willkürlich“ auf einen bestimmten Wert wie z.B.  $R^2 = 0.10$  fest, so hat sich eine psychologische Hypothese, die die statistische Alternativhypothese impliziert, in einem Forschungsprogramm dann bewährt, wenn  $Q = \alpha^x (1 - \alpha)^{m-x} / (1 - \beta)^x \beta^{m-x} < 1$ , wobei  $m$  die Anzahl der Untersuchungen und  $x$  die Anzahl signifikanter Resultate ist. Aber auch bei  $Q > 1$  könnte behauptet werden, daß die psychologische Hypothese gestützt ist; da von einem Minimaleffekt  $R^2 = 0.10$  ausgegangen wurde, der sich als zu groß herausgestellt habe, müsse  $\beta$  tatsächlich viel größer als festgelegt gewesen sein, so daß etwa 3 von 10 signifikanten Ergebnissen die Hypothese eines kleineren Effekts stützen. Eine derartige Behauptung müßte begründet werden können (z.B. zu geringe Durchschlagskraft der unabhängigen Variablen), und nur im Falle einer Begründung (und nicht einer beliebigen Exhaustion „negativer“ Befunde) wäre eine prüfbare Hypothese generiert worden. An dieser Stelle zeigt sich, daß mit der skizzierten Vorgehensweise Falsifikationen logisch und empirisch nicht erzwungen werden können. Der Forscher kann sich aufgrund in diesem Sinne geplanter Untersuchungen nur zur Falsifikation und zum Neuentwurf von Hypothesen *entschließen*. Die Planung der Untersuchungen ist so anzulegen, daß ein derartiger Entschluß überhaupt ermöglicht wird. Eine Planung, wie sie hier skizziert wurde (ausführlich dazu Bredenkamp, 1980), würde vermutlich auch erreichen, daß sowohl seitens des Forschers wie der Herausgeber von Fachzeitschriften „negative“ Resultate veröffentlicht werden, so daß die häufig beklagte Kumulation des statistischen  $\alpha$ -Fehlers in den Publikationen vermieden werden könnte.

Gegen die skizzierte Vorgehensweise sind durch Deppe (1977) und Glaser (1979) Bedenken angemeldet worden. Deppe (1977) weist, ganz im Sinne der obigen Ausführungen, darauf hin, daß ein Modell durch das Ergebnis eines Signifikanztests logisch nicht zu widerlegen ist. Andererseits schreibt er: „Wenn in *vielen* Experimenten zum Begriffserwerb z.B. ein negativ beschleunigter Abfall der Fehlerkurven beobachtet wird, und wenn ein Modell hier einen konstanten Abfall der Fehler voraussagt, ist eher anzunehmen, daß das Modell systematisch falsche Annahmen macht, als wenn die Fehlerkurve nur einmal beobachtet wurde“ (Deppe, 1977, 166). über derartige Abweichungen der tatsächlichen von den prognostizierten Resultaten kann man mit Hilfe eines Signifikanztests entscheiden, und wenn aus mehreren Untersuchungen die dem Modell widersprechenden Ergebnisse der Tests konvergieren, kann man sich zu einer Falsifikation und Abänderung des Modells entschließen, vorausgesetzt, die Tests wurden im Sinne obiger Ausführungen geplant. Die Ausführungen Deppes scheinen uns der dargestellten Konzeption des Hypothesentests nicht zu widersprechen, wenn anerkannt wird, daß die Falsifikation nur aufgrund eines methodologischen Beschlusses möglich ist, z.B.: „Akzeptiere bei kleinen Fehlerwahrscheinlichkeiten  $H_0$  als eine der psychologischen Hypothese widersprechende Populationsaussage“ (vgl. Bredenkamp, 1980).

Glaser (1979, 135) anerkennt das skizzierte Verfahren als sinnvoll, betont jedoch, daß es nicht der Falsifikation von psychologischen Hypothesen dienen könne. Nur deterministisch formulierte Hypothesen sind falsifizierbar, nicht jedoch statistische Hypothesen, da sie kein Ereignis verbieten. Diese Feststellungen treffen selbstverständlich zu und stimmen mit der Aussage Deppes überein, daß das Ergebnis eines Signifikanztests logisch ein Modell nicht widerlegen könne. Glaser (1979) scheint zwar die Möglichkeit zu akzeptieren, aufgrund eines statistischen Tests eine vorgeordnete psychologische Hypothese zu verwerfen - von Falsifikation sollte s.E. nicht gesprochen werden -, wähnt sich aber auf sichererem Boden, wenn psychologische Gesetze von vornherein als Wahrscheinlichkeitsaussagen formuliert werden, die nicht falsifiziert werden können. Dem ist jedoch entgegenzuhalten, daß die de facto bei der Durchführung statistischer Tests geprüften statistischen Hypothesen selten mit den psychologischen identisch sind. Beispiele für diese Behauptung finden sich etwa bei Bredenkamp (1972, 1980) und bei Hager und Westermann in diesem Band. Psychologische Hypothesen beziehen sich auf individuelles Geschehen; die in der empirischen Psychologie üblichen statistischen Tests prüfen aber Populationsaussagen über die Gleichheit oder Verschiedenheit von Parametern. Will man diese Prüfungen nicht von vornherein als sinnlos betrachten, muß ein Bezug zwischen psychologischen Hypothesen und Populationsaussagen hergestellt werden. Deterministisch formulierte psychologische Hypothesen implizieren Populationsaussagen. Auf erstere richtet sich der Falsifikationsanspruch. Falsifikationsinstanzen sind statistische Populationsaussagen, über deren Zutreffen entschieden werden muß. Dabei ist zu gewährleisten, daß die mit dieser Entscheidung verbundenen Fehlermöglichkeiten gering sind, und genau dazu dient das skizzierte Verfahren. Wenn Glaser (1979, S. 125) sagt, daß „eine deterministische psychologische Hypothese . . . mit einem widersprechenden Datum von einer Person, die dem Individuumbereich der Hypothese angehört, erledigt“ ist, scheint er vorauszusetzen, daß immer für einzelne Personen singuläre Existenzsätze formuliert werden können, die im Einklang oder im Widerspruch zur Hypothese stehen. Dies aber ist in der Psychologie eben häufig nicht der Fall (vgl. Bredenkamp 1972, 1980), und dennoch sind die zu prüfenden Hypothesen vielfach deterministisch formuliert. Die implikative Verknüpfung zwischen psychologischer Hypothese und Populationsaussage und daraus resultierende Veränderungen am in der Psychologie üblichen statistischen Test scheinen uns der einzige Weg zu sein, dieser Situation gerecht zu werden. Freilich können dann nur bei Vereinbarung methodologischer Regeln Hypothesen falsifiziert werden. Dies aber ist keine Besonderheit des Forschungsprozesses, die erst durch das statistische Hypothesentesten ins Spiel kommt. Popper (1966) hat in seiner „Logik der Forschung“ immer wieder darauf hingewiesen, daß auch singuläre Existenzsätze Dispositionsprädikate enthalten, die nicht vollständig auf beobachtbare Gegebenheiten zurückgeführt werden können. Die Basissätze müssen deshalb innerhalb der Falsifikationstheorie Poppers (1966) durch Festsetzung aner-

kannt werden. Übernehmen statistische Populationsaussagen die Funktion von Basissätzen, so müssen methodologische Festsetzungen vereinbart werden, die ihre Anerkennung ermöglichen. Die von Glaser (1979) akzeptierte Modifikation des statistischen Hypothesentestens scheint uns in der Psychologie überhaupt nur unter dieser Zielsetzung begründbar zu sein. Bei der pfadanalytischen Überprüfung linearer Kausalstrukturen verhält es sich nicht anders. Auch diese Modelle sind im Grunde deterministisch formuliert. Durch die Hereinnahme der impliziten Variablen wird behauptet, daß die Variation der endogenen Variablen völlig aufgeklärt werden kann. Da diese Variablen unbekannt sind, lassen sich aus dem Modell jedoch nur Korrelationen zwischen den expliziten Variablen ableiten (und nicht etwa linear funktionale Beziehungen), und aufgrund dieser Korrelationen wird das Modell überprüft.

Es sei hier noch kurz auf eine erweiterte Theorie des statistischen Hypothesentestens eingegangen, die Witte (1980) formuliert hat. Danach vollzieht sich die Beurteilung von statistischen Hypothesen in folgenden Schritten:

- (1)  $\alpha$  und  $\beta$  werden klein gewählt; es wird eine Effektgröße, die mit der Wahrscheinlichkeit  $1 - \beta$  entdeckt werden soll, festgelegt, und der benötigte Stichprobenumfang wird bestimmt. Dieses Vorgehen entspricht den bisherigen Ausführungen.
- (2) Anstelle eines Signifikanztests wird ein likelihood-Test (vgl. dazu Wendt in diesem Band) für zwei einfache Hypothesen durchgeführt, dem der Vorzug gegenüber dem Signifikanztest gegeben wird, weil er nur das eingetretene Ergebnis (etwa  $\bar{X}=2$ ) und nicht auch größere Abweichungen vom Parameter  $\mu_0=0$  unter  $H_0$  verarbeitet. Entspricht die Effektstärke etwa einem Parameter  $\mu_1=5$ , so lautet der Test:

$$\psi = \frac{L(\mu_0=0/\bar{X}=2)}{L(\mu_1=5/\bar{X}=2)}$$

Der Wert für  $\psi$  muß die Grenze  $g_1 = \frac{1-\alpha}{\beta}$  überschreiten, wenn  $H_0$ , oder  $g_2 = \frac{\alpha}{1-\beta}$  unterschreiten, wenn  $H_1$  besser gestützt ist, wobei diese Grenzen der Theorie sequentieller Verfahren von Wald (vgl. dazu Wendt in diesem Band) entnommen wurden. Der Prüfschritt ist nur dann befriedigend ausgefallen, wenn  $g_1$  über- oder  $g_2$  unterschritten wird.

- (3) Sind die beiden ersten Prüfschritte positiv ausgegangen, dann wird ermittelt, ob die größere Stützung der einen Hypothese auf einem hohen likelihood-Wert der akzeptierten oder auf einem geringen likelihood-Wert der abgelehnten Hypothese beruht, indem die likelihoods an der maximalen likelihood gemessen werden. Dabei muß der kritische Quotient  $Q_C = 1 - \sqrt{\alpha(1-\beta)}$  durch die in Schritt 2 akzeptierte Hypothese erreicht oder überschritten werden.
- (4) Schließlich ist die Größe des Effekts zu schätzen, wobei Witte (1980) als



Faustregel angibt, daß nicht weniger als 10% der totalen Varianz aufgeklärt sein sollten.

Die besser gestützte Hypothese wird akzeptiert, wenn alle Prüfschritte positiv ausgefallen sind; nur dann wird auch die weniger gestützte Hypothese abgelehnt. Falls wenigstens einer der vier Prüfschritte negativ ausfällt, wird keine Entscheidung getroffen.

Kritisch ist zu diesem Vorgehen anzumerken, daß die Prüfschritte 2 und 3 negativ ausgehen können, wenn ein sehr großer Effekt besteht, der vierte Prüfschritt also positiv ausgefallen ist. Dies wäre etwa der Fall, wenn für normal verteilte Daten die Hypothesen  $\mu_0=0$  und  $\mu_1=1$  bei  $\sigma=1$  getestet werden,  $\bar{X}$  aber 10 oder größer ist. Diese Schwierigkeit hängt damit zusammen, daß nur einfache statistische Hypothesen getestet werden. Selten jedoch liegt in der Psychologie dieser Fall vor.

Bei den besprochenen Konzeptionen fand das Bayessche Hypothesentesten keine Berücksichtigung, das Wendt in diesem Band ausführlich darstellt und favorisiert (zur Kritik an der Verwendung des Bayesschen Theorems beim statistischen *Hypothesentesten*, vgl. Rützel, 1979, 1980). Die von Vertretern der Bayes-Statistik oft behauptete und auch aufgezeigte Voreingenommenheit von Signifikanztests gegen die Nullhypothese kann, wenn man einmal die Prämisse der Bayesianer, daß Verteilungsparameter eine Zufallsvariable sind, der Wahrscheinlichkeiten (oder Wahrscheinlichkeitsdichten) zuzuordnen sind, akzeptiert, praktisch durch die eingangs beschriebene Kontrolle der Wahrscheinlichkeiten  $\alpha$  und  $\beta$  und deren Identifikation mit kleinen Werten aufgehoben werden (vgl. dazu Bredenkamp, 1972). Allerdings kann man das Bayessche Theorem dann gut verwenden, wenn Hypothesen im Lichte ganzer Forschungsprogramme beurteilt werden und ermittelt werden soll, welche der statistischen Hypothesen besser gestützt ist. Unterteilt man den Parameterraum in zwei Teilklassen  $H_0$  und  $H_1$ , die zu Beginn eines Forschungsprogramms beide für gleich wahrscheinlich gehalten werden, und zerlegt man die Menge aller möglichen Ergebnisse ebenfalls in zwei Teilklassen „signifikant“ und „insignifikant“, so ist bei Verwendung des Bayesschen Theorems immer die Wahrscheinlichkeit für  $H_0$  im Lichte der Daten ( $P(H_0/D)$ ) größer als die für  $H_1$ , wenn der zuvor eingeführte Wert  $Q = \alpha^x(1-\alpha)^{m-x}/(1-\beta)^x\beta^{m-x}$  größer als 1 ist, da  $P(H_0/D)=Q/(Q+1)$ . Unter den genannten Voraussetzungen ist  $Q$  ein Bayessches Stützmaß; hinzu kommt, daß  $\alpha$  und  $\beta$ , wie besprochen, kontrolliert wurden.

### 3. Dynamische Modelle

Die Zeit und damit korrelierte Veränderungen der Variablen wurden bisher nicht berücksichtigt. Die in Abschnitt 1 besprochenen Modelle eignen sich

nach Hummell und Ziegler (1976) dann zur Darstellung kausaler Prozesse, wenn eine der folgenden Bedingungen erfüllt ist:

- (1) Auf eine Änderung der exogenen Faktoren erfolgt sofort eine Änderung der endogenen Variablen.
- (2) Zum Zeitpunkt der Beobachtung befindet sich der Prozeß in einem Gleichgewichtszustand: Die Werte aller Variablen ändern sich nicht mehr.
- (3) Es liegt ein Prozeß zugrunde, der die Korrelationen einem asymptotischen Grenzwert zustreben läßt, der zum Zeitpunkt der Beobachtung annähernd erreicht ist.

Die erste Annahme wird man häufig in der experimentellen Psychologie treffen müssen, wenn nämlich sofort nach der Manipulation einer unabhängigen Variablen die abhängige Variable erhoben wird. Treten die Effekte jedoch zeitverzögert auf, so würde das geprüfte experimentelle Kausalmodell fälschlich falsifiziert werden. Sinnvoll ist die erste Annahme wohl nur als zutreffend zu unterstellen, wenn ausdrücklich die sofortige Gedächtnisleistung, die sofort eintretenden Stimmungsveränderungen usw. analysiert werden sollen, wenn also der Validitätsanspruch für endogene Variablen von vornherein ausdrücklich auf den Zeitpunkt kurz nach der Manipulation der unabhängigen Variablen eingengt wird. Anderenfalls muß die Bedingung (2) oder (3) erfüllt sein. Allerdings läßt sich deren Zutreffen wohl selten überprüfen. In diesem Fall ist es ratsam, zu dynamischen Modellen überzugehen, die den Zeitparameter berücksichtigen und die endogenen Variablen zu mehreren Zeitpunkten enthalten.

Die Konzeption dynamischer Modelle führt insofern zu Problemen bei der Schätzung der Pfadkoeffizienten, als die impliziten Faktoren, welche die endogenen Variablen beeinflussen, nicht mehr als unkorreliert angesehen werden können. In Abb. 7 ist ein Modell mit einer exogenen Variablen zum Zeitpunkt  $t-1$  (die Zeit wird als diskrete Variable behandelt) und zwei endogenen Variablen aufgeführt, die zu verschiedenen Zeitpunkten erhoben wurden. Auf den ersten Blick unterscheidet sich dieses Modell von dem in Abb. 3a nicht. Es besteht jedoch ein Unterschied: Da die endogene Variable zu verschiedenen Zeitpunkten erhoben wurde, wird eine Korrelation (sog. Autokorrelation) zwischen den impliziten Faktoren  $U_{t-1}$  und  $U_t$  angenommen. Bei mehr als zwei Meßzeitpunkten können auch Autokorrelationen höherer Ordnung entstehen (z.B. zwischen  $U_{t-2}$  und  $U_t$ ). Wenn man nicht davon ausgehen kann, daß die Autokorrelationen Null sind, lassen sich die Pfadkoeffizienten auch nicht mehr so, wie bisher dargestellt, schätzen. Auf Modelle, die Autokorrelation einbeziehen, gehen Möbus und Nagl, die auch andere Möglichkeiten der Verlaufsanalyse untersuchen, in diesem Band ausführlich ein (vgl. auch Schubö et al. in Band 4 dieser Enzyklopädie); Verteilungsfreie Analysen von Zeitreihen behandelt ausführlich Lienert (1978)).

In der experimentellen Psychologie werden oftmals Veränderungen erfaßt, z.B. bei der Analyse von Lernkurven. Meistens wird eine varianzanalytische Auswertung vorgenommen, und es muß u.a. unterstellt werden, daß die Korrelationen zwischen den

Werten der abhängigen Variablen zu verschiedenen Meßzeitpunkten gleich groß sind. Verletzungen dieser Annahmen führen dazu, daß der F-Test zu häufig eine richtige Nullhypothese zurückweist. Diesen Fehler versucht man dadurch zu vermeiden, daß die Anzahl der Freiheitsgrade entsprechend des Ausmaßes der Heterogenität der Korrelationen zwischen den Meßwerten reduziert wird (vgl. dazu Huyng, 1978). Eine andere Möglichkeit besteht darin, die abhängigen Variablen nach bestimmten Regeln zu transformieren und multivariate Hypothesentests durchzuführen, die alle Hypothesen unter weniger restriktiven Annahmen zu prüfen gestatten, welche bei der Erfüllung der Voraussetzungen mit einer univariaten Varianzanalyse ebenfalls prüfbar wären. Es lassen sich also etwa auch Interaktionen zwischen einer oder mehreren Behandlungsvariablen mit dem Zeitfaktor auf Signifikanz prüfen. Eine ausgezeichnete Darstellung des Vorgehens findet sich bei McCall und Appelbaum (1973).

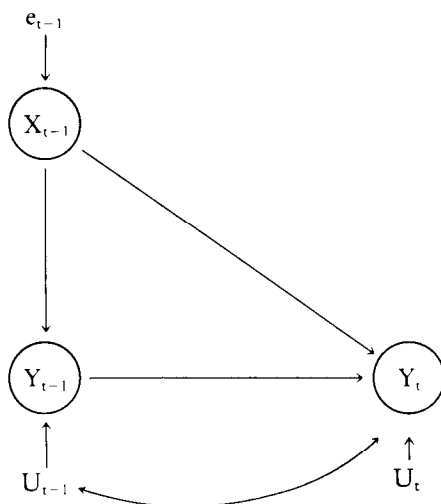


Abb. 7:

In dem Band der Enzyklopädie, dem dieses einführende Kapitel vorangestellt ist, findet sich schließlich eine Einführung in die Computer-Simulation psychischer Prozesse von Ueckert. Simulationsmodelle sind, wie der Autor betont, in der Regel als dynamische Modelle konzipiert. Ihre Besonderheit liegt darin, daß eine Theorie über den Gegenstandsbereich in eine Sprache übersetzt werden muß, die der Rechner versteht, so daß er „in allen auftretenden Situationen theoriegemäßes Verhalten zeigen kann. Dann kann das gleiche Experiment mit menschlichen und der „künstlichen“ Versuchsperson durchgeführt werden, und die Ergebnisprotokolle können miteinander verglichen werden“ (Deppe, 1977, 126).

Ueckert hebt die verschiedenen Datenquellen hervor, die für die Konzeption eines Simulationsmodells herangezogen werden können; dem Experiment komme wegen seiner methodologisch begründeten Künstlichkeit nur geringe Bedeutung zu. Hierin erblicken wir allerdings keine Besonderheit von Simulationsmodellen. Für die Generierung von experimentell zu prüfenden Hypothesen kommen auch alle möglichen Quellen in Frage, u.U. z.B. auch das einführende Verstehen (vgl. dazu Patzig, 1973). Die experimentelle Kontrolle wird erst bei der Prüfung der wie auch immer gefundenen Hypothese wichtig, um fälschliche Bestätigungen oder Widerlegungen zu vermeiden. Wie das obige Zitat von Deppe verdeutlicht, muß auch das Simulationsmodell überprüft werden, und wenn es sich auf menschliches Verhalten bezieht, so geschieht diese Prüfung durch den Vergleich des „Verhaltens“ des Computers mit dem Verhalten von Menschen. Deppe (1977) zeigt, wie Simulationsmodelle der Konzeptidentifikation aufgrund derartiger Vergleiche abgeändert wurden. Nach Ueckert sind Simulationsmodelle nicht falsifizierbar, da sie aus einer logischen und einer empirischen Komponente bestehen. Der logische Kern bleibt unabhängig davon, welche empirischen Daten auftreten, unbeeinflußt. Die empirische Komponente beinhaltet die Menge der intendierten Anwendungen eines Modells; deren Prüfung kann nur ergeben, ob ein Modell in einer bestimmten Situation anwendbar ist oder nicht. Dieser ursprünglich für hoch entwickelte physikalische Theorien formulierte sog. non-statement view von Sneed kann wohl nur mit Einschränkungen auf Theoriebildungen in der Psychologie übertragen werden (vgl. dazu Herrmann, 1976). Die Nicht-Anwendbarkeits-Interpretation erscheint für elaborierte und gut bewährte psychologische Theorien sinnvoll, wenn nicht eine andere Theorie vorhanden ist, deren Anwendbarkeitsbereich den des konkurrierenden Modells enthält, zusätzlich aber noch andere Anwendungen zuläßt. Liegt also der Fall vor, daß in diesem Sinne etwa die Theorie der Konzeptidentifikation von Levine (1975, Kap. 11 und 12) mit der von Bower und Trabasso (1964) verglichen werden kann, so wäre aufgrund der Falsifikationstheorie die letztgenannte als falsifiziert zu bezeichnen, während die von Levine als bewährt gelten kann. Durch die Formulierung beider Theorien in einer Sprache, die der Computer versteht, ändert sich hieran nichts.

Deppe (1977), der zu einer sehr ausgewogenen Beurteilung beim Vergleich von Simulationsmodellen mit mathematischen Modellen in der Psychologie gelangt, stellt zunächst heraus, daß beide eine psychologische Theorie voraussetzen. Sie sind kein Ersatz, sondern Hilfsmittel für die Theorienbildung. Wenn viele Größen zueinander in Beziehung stehen, können Simulationsmodelle die Komplexität häufiger als mathematische Modelle angemessen berücksichtigen; dadurch wird eine ganzheitliche Betrachtung wiederbelebt. Andererseits kann man aus einem mathematischen Modell exakte Prognosen für die *Population* der Menschen herleiten, die gemäß den Modellannahmen „funktionieren“. Dagegen sind Prognosen eines Simulationsmodells eine Stichprobe, die in un-

bekanntem Ausmaß verzerrt ist: „Wenn die Daten deutlich voneinander abweichen, bietet sich . . . die zusätzliche Möglichkeit, dies unter Hinweis auf den Stichprobencharakter der Prognosen zu entkräften“ (Deppe, 1977, 144). Mit anderen Worten: Mathematische Modelle sind strenger prüfbar als Simulationsmodelle, und dieser Gesichtspunkt, im Rahmen der Falsifikationstheorie betrachtet, ist sehr wichtig (vgl. dazu Hager und Westermann in diesem Band). Andererseits setzt dieses Kriterium voraus, daß die mathematische Modellbildung mit der Simulation verglichen werden kann, und wegen der besseren Handhabbarkeit komplexer Beziehungen durch Simulationsmodelle ist dieser Vergleich nicht immer möglich.

## *Literatur*

- Bower, G. H. & Trabasso, T. 1964. Concept identification. In: Atkinson, R. C. (Ed.): Studies in mathematical psychology. Stanford: Stanford University Press.
- Brandstädter, J. 1976. Soziale Schicht, Umwelt und Intelligenz: Eine Pfadanalyse der Korrelationsbefunde von Marjoribanks. Psychologische Beiträge, 18, 35-53.
- Brandstädter, J. & Bernitzke, F. 1976. Zur Technik der Pfadanalyse. Ein Beitrag zum Problem der nichtexperimentellen Konstruktion von Kausalmodellen. Psychologische Beiträge, 18, 12-34.
- Bredenkamp, J. 1969. über die Anwendung von Signifikanztests bei theorietestenden Experimenten. Psychologische Beiträge, 11, 275-285.
- Bredenkamp, J. 1972. Der Signifikanztest in der psychologischen Forschung. Frankfurt a. M.: Akademische Verlagsgesellschaft.
- Bredenkamp, J. 1980. Theorie und Planung psychologischer Experimente. Darmstadt: Steinkopff.
- Bredenkamp, J. 1982. Verfahren zur Ermittlung des Typs einer statistischen Interaktion. Psychologische Beiträge (im Druck).
- Campbell, D. T. & Stanley, J. C. 1973. Experimental and quasi-experimental designs for research on teaching. In: Gage, N. L. (Ed.): Handbook of research on teaching. Chicago: Rand McNally.
- Cohen, J. 1977<sup>2</sup>. Statistical power analysis for the behavioral sciences. New York: Academic Press.
- Deppe, W. 1977. Formale Modelle in der Psychologie. Stuttgart: Kohlhammer.
- Gaensslen, H. & Schubö, W. 1973. Einfache und komplexe statistische Analyse. München: Reinhardt.
- Glaser, W. R. 1979. Statistische Entscheidungsverfahren über Hypothesen in den Sozialwissenschaften. In: Albert, H. & Stapf, K. H. (Ed.): Theorie und Erfahrung. Stuttgart: Klett-Cotta.
- Herrmann, T. 1976. Die Psychologie und ihre Forschungsprogramme. Göttingen: Hogrefe.

- Hummell, H. J. & Ziegler, R. 1976. Zur Verwendung linearer Modelle bei der Kausalanalyse nicht-experimenteller Daten. In: Hummell, H. J. & Ziegler, R. (Ed.): Korrelation und Kausalität, Band 1, E5-E137. Stuttgart: Enke.
- Huyngh, H. 1978. Some approximate tests for repeated measurement designs. *Psychometrika*, 43, 161-175.
- Kerlinger, F. N. & Pedhazur, E. J. 1973. Multiple regression in behavioral research. New York: Holt, Rinehart and Winston.
- Lehmann, G. 1980. Nichtlineare „Kausal-“ bzw. Dominanz-Analysen in Psychologischen Variablensystemen. *Zeitschrift für experimentelle und angewandte Psychologie*, 27, 257-276.
- Levine, M. 1975. A cognitive theory of learning. Research on hypothesis testing. Hillsdale: Erlbaum.
- Lienert, G. A. 1978. Verteilungsfreie Methoden in der Biostatistik. Band II. Meisenheim am Glan: Anton Hain.
- McCall, R. B. & Appelbaum, M. I. 1973. Bias in the analysis of repeated-measures designs: Some alternative approaches. *Child Development*, 44, 401-415.
- Patzig, G. 1973. Erklärung und Verstehen. Bemerkungen zum Verhältnis von Natur- und Geisteswissenschaften. *Neue Rundschau*, 3, 392-413.
- Popper, K. R. 1966<sup>2</sup>. Logik der Forschung. Tübingen: Mohr.
- Rützel, E. 1979. Bayessches Hypothesentesten und warum die Bayesianer Bias-ianer heißen sollten. *Archiv für Psychologie*, 131, 211-232.
- Rützel, E. 1980. Korrektur zu E. Rützel: Bayessches Hypothesentesten und warum die Bayesianer Bias-ianer heißen sollten (Bayesianism or Biasianism?). *Archiv für Psychologie*, 132, 187-188.
- Wippich, W. & Bredenkamp, J. 1979. Bildhaftigkeit und Lernen. Darmstadt: Steinkopff.
- Witte, E. H. 1980. Signifikanztest und statistische Inferenz. Stuttgart: Enke.

## 2. Kapitel

# Planung und Auswertung von Experimenten\*)

*Willi Hager und Rainer Westermann*

### *Vorbemerkungen*

Ein wesentliches Ziel jeder empirischen Wissenschaft besteht darin, zu fundierten Kausalaussagen zu gelangen, also zu Aussagen über die Ursachen oder Bedingungen der jeweils interessierenden beobachtbaren Sachverhalte. Die bevorzugte Methode zur Überprüfung entsprechender Vermutungen oder *Kausalhypothesen* ist das Experiment, auch in der Psychologie und in benachbarten Sozial- und Verhaltenswissenschaften (Abschn. 1.2). Das Ausmaß, in dem ein bestimmtes Experiment zur Prüfung einer solchen Kausalhypothese geeignet ist, werden wir als dessen *Validität* bezeichnen (Abschn. 1.3).

Soll im Experiment eine Kausalhypothese geprüft werden, müssen die in ihr vorkommenden (theoretischen) Begriffe in beobachtbare Variablen „übersetzt“ werden. Fehler bei dieser „Operationalisierung“ beeinträchtigen die „*Variablenvalidität*“ der Untersuchung (Teil 2). Ob mit Hilfe eines Experiments überhaupt Aussagen über Ursachen möglich sind, hängt von seiner „*internen Validität*“ ab (Teil 3). Inwieweit ein Experiment eine Prüfung für die betrachtete Kausalhypothese ist, wird auch dadurch beeinflusst, mit welchen Personen und in welcher Situation es durchgeführt wird. Diese in ihrer Bedeutung für die wissenschaftliche Theorienbildung häufig unterschätzte „*Populations- und Situationsvalidität*“ wird im Teil 4 besprochen. Die bisher erwähnten Aspekte der experimentellen Validität können sich in verschiedener Weise gegenseitig beeinflussen, und zwar je nach Art der geprüften Hypothese fördernd oder hemmend (Teil 5). Im Teil 6 zeigen wir, daß über die Gültigkeit einer wissenschaftlichen Hypothese entschieden werden kann, indem über die

---

\*) Wir danken den Kollegen Marcus Hasselhorn (Heidelberg), Falk Leichsenring (Göttingen) und Werner Wippich (Trier) für ihre kritischen Anmerkungen zu einigen früheren Fassungen dieser Arbeit und Frau Gabriele Reimann für das Anfertigen der Zeichnungen.

Gültigkeit von aus ihr abgeleiteten *statistischen Hypothesen* entschieden wird. Diese Entscheidung erfolgt mit Hilfe von *Signifikanztests* (Teil 7). Die wichtigsten Fehler, die dabei gemacht werden können und die dann die „*statistische Validität*“ des Experiments herabsetzen, besprechen wir im Teil 8. Dabei werden sich wesentliche Hinweise für die Auswahl derjenigen Versuchspläne und Auswertungsmethoden ergeben, die für ein bestimmtes Experiment wahrscheinlich am besten geeignet sind. Auf zwei Aspekte der statistischen Validität gehen wir in den Teilen 9 und 10 besonders ein: auf Maße für die Größe des „experimentellen Effekts“ und auf die begründete Wahl des Stichprobenumfangs. Eine auf diesen Überlegungen basierende Planungs- und Entscheidungsstrategie stellen wir im Teil 11 dar.

Obwohl wir detailliert auf bestimmte Fragen und Probleme der statistischen Auswertung eingehen müssen, setzen wir die Kenntnis der gebräuchlichen Testverfahren voraus bzw. geben lediglich an, wo man sie sich verschaffen kann.

Im Laufe unserer Erörterungen werden wir immer wieder auf die verbreiteten sog. „Standard-Versuchspläne“ stoßen. Einige ihrer Vor- und Nachteile werden wir jeweils kurz vor dem Hintergrund der verschiedenen Aspekte der experimentellen Validität - dem zentralen Anliegen unseres Artikels - diskutieren; darüber hinausgehende Hinweise zur Anwendung und Auswertung dieser Pläne können dann jeweils den detaillierten Literaturangaben entnommen werden. Überhaupt sollen die zahlreichen Verweise den Leser anregen, sich mit einigen Originalarbeiten zu ihn interessierenden Problemen zu beschäftigen. Er wird dann feststellen, daß viele der hier angesprochenen Probleme noch kontrovers diskutiert werden, so daß man in zahlreichen Einzelfällen sicher auch andere Meinungen und Empfehlungen vertreten kann als die beiden Autoren dieser Arbeit. Insbesondere kann man zu anderen Ergebnissen kommen, wenn man einen anderen wissenschaftstheoretischen Ausgangspunkt wählt.

Insgesamt kommt es uns vornehmlich darauf an, dem „Produzenten“ experimenteller Ergebnisse konkrete Hinweise zu vermitteln, wie er sein Experiment so planen, durchführen und auswerten kann, daß die Resultate zur Entscheidung über die Gültigkeit seiner wissenschaftlichen Hypothese beitragen können.

Der „Konsument“ experimenteller Befunde soll angeregt werden, sich kritisch mit Ansatz, Durchführung und Auswertung ihn interessierender Untersuchungen zu befassen.



# 1. Einleitung

## 1.1 Einige Begriffsbestimmungen

Faßt man die wesentlichen Gemeinsamkeiten der zahlreichen Definitionen des Experiments (siehe z.B. Bredenkamp, 1969a; Zimmermann, 1972; Fietkau, 1973) zusammen, gelangt man etwa zu folgenden Kennzeichen:

- (1) Im Experiment werden bestimmte Bedingungen willkürlich (absichtlich) hergestellt.
- (2) Die hergestellten Bedingungen werden in bestimmter Weise (systematisch) variiert.
- (3) Durch den Vergleich verschiedener Beobachtungen wird der Einfluß dieser Variation auf bestimmte andere Merkmale festgestellt.
- (4) Alle anderen Bedingungen versucht man konstant zu halten oder zu kontrollieren.

Die vom Experimentator (abgekürzt: E) variierten Bedingungen bezeichnet man als die *unabhängigen Variablen* (UVn) oder *Faktoren* des Experiments. Im einfachsten Fall liegt eine unabhängige Variable (UV) mit zwei Ausprägungen vor. Diese Ausprägungen werden auch *Modalitäten*, Stufen oder (Behandlungs- bzw. Treatment-) Bedingungen genannt, oft auch nur kurz Treatments. Als Beispiel denke man an die UV „Art der Verstärkung“ (symbolisiert als A) mit den Modalitäten „Belohnung“ (abgekürzt als  $A_1$ ) und „Bestrafung“ ( $A_2$ ); allgemein bezeichnen wir die Modalitäten mit  $A_j$ , wobei gilt:  $j = 1, 2, \dots, l, l', \dots > J$ .

In der Psychologie sind die Modalitäten der UV i.a. wie im angeführten Beispiel qualitativ gestuft; nur in Ausnahmefällen besteht die Möglichkeit, eine UV mit quantitativen Abstufungen einzuführen (als Beispiel denke man an die Darbietungszeiten in einem Lernexperiment).

Werden unabhängig voneinander mehrere Merkmale variiert, spricht man von einem „*mehrfaktoriellen* (oder multidimensionalen) Experiment“. Als Beispiel für eine zweite UV sei die „Häufigkeit der Verstärkung“ (symbolisiert als B) genannt, und zwar mit den Ausprägungen „Nach jeder Reaktion“ ( $B_1$ ) und „Nach jeder 3. Reaktion“ ( $B_a$ ). Allgemein benennen wir die Modalitäten des Faktors B mit  $B_k$ , wobei gilt:  $k = 1, 2, \dots, m, m', \dots, K$ . In mehrfaktoriellen Experimenten wird i.a. jede Bedingung jeder UV mit jeder Bedingung jeder anderen UV kombiniert („gekreuzt“); vgl. jedoch Abschnitt 8.3.5.

Die Sachverhalte, die als Auswirkungen der vorgenommenen Bedingungsvariation betrachtet werden, stellen die *abhängigen Variablen* (AVn) des Experiments dar, z.B. die „Anzahl der verbalen Aggressionen gegen den Versuchsleiter“ ( $Y_u$ ) oder die „Anzahl der Durchgänge bis zum Erreichen des Lernkri-

teriums“ ( $Y_v$ ). Untersuchungen, in denen die Werte auf nur einer AV erhoben werden, heißen *univariat*, solche mit mehreren *multivariat*.

Diese Begriffsbestimmungen sind nur als erste Orientierung gedacht, Präzisierungen ergeben sich aus dem weiteren Inhalt dieser Arbeit. Insbesondere werden wir im Abschnitt 3.6 zu einer genaueren Definition des Experiments gelangen, die auch eine Abgrenzung gegenüber anderen Forschungsmethoden erlaubt.

## 1.2 Das Experiment als Methode zur Prüfung von Kausalaussagen

Um beschreiben zu können, wann Planung und Auswertung eines Experiments adäquat sind und welche Fehler dabei gemacht werden können, wollen wir von Ziel und Zweck des Experiments ausgehen: In der Fachliteratur wird das Experiment als die einzige Methode bezeichnet, die generell zur Überprüfung von Kausalhypothesen geeignet ist (Siebel, 1965; Aronson & Carlsmith, 1968; Miller, 1970; Gadenne, 1976; Cook & Campbell, 1979). Kausalhypothesen und alle anderen Arten von Vermutungen, vorläufigen Problemlösungen, theoretischen Ableitungen, Vorhersagen usw., die empirischen Untersuchungen vorgeordnet sind, werden wir zusammenfassend „wissenschaftliche Hypothesen“ (WH) nennen. Zur Verdeutlichung wollen wir von einer fiktiven Problemstellung ausgehen, die uns als Beispiel auch bei unseren weiteren Erörterungen zuweilen begegnen wird:

Da man befürchtet, daß durch die mangelnde Integration von Ausländern (Gastarbeitern) schwere soziale Konflikte auftreten, sei die politische Entscheidung gefallen, die Benachteiligung von Ausländern zu beenden und ihre Eingliederung in die Gesellschaft zu fördern. Da dieser Beschluß nicht einhellig von der Bevölkerung unterstützt wird, soll als einer der ersten Schritte ein Programm zur Veränderung der negativen Einstellung gegenüber Ausländern durchgeführt werden.

Wir wollen uns vorstellen, ein Team von Psychologen und Sozialwissenschaftlern habe den Auftrag bekommen, dieses Programm zu planen, durchzuführen und seine Wirksamkeit zu überprüfen.

In der Fachliteratur ist eine große Zahl von theoretischen Vorstellungen darüber zu finden, durch welche Maßnahmen Einstellungen verändert werden können. Eine der bekanntesten Theorien in diesem Zusammenhang ist die kognitive Dissonanztheorie von Festinger (1978; vgl. Irle, 1975; Frey, 1978).

Aus der Theorie der kognitiven Dissonanz läßt sich unter anderem die folgende Hypothese ableiten: „Wenn zwischen der Einstellung einer Person zu einem bestimmten Objekt und einem anderen kognitiven Element eine Dissonanz besteht, dann verändert sich die Einstellung so, daß diese Dissonanz

vermindert wird.“ Hier wird die Erzeugung von Dissonanz (D) als hinreichende Bedingung für die Änderung der Einstellung bezeichnet. Wir wollen diese spezielle wissenschaftliche Hypothese als WH, bezeichnen und durch „ $D \mapsto dE$ “ symbolisieren. Dabei soll  $dE$  eine Veränderung der betrachteten Einstellung bezeichnen. Falls die Hypothese gültig ist, falls sie also eine psychologische Gesetzmäßigkeit ausdrückt, gilt folgendes: Treten die Ereignisse „D“ und „ $dE$ “ ein, kann „D“ als die *Ursache* von „ $dE$ “ bezeichnet werden (Gadenne, 1976, 29). Von daher können wir unsere obige Vermutung als Kausalhypothese bezeichnen.<sup>1)</sup>

Trotz einer oberflächlichen Ähnlichkeit wäre es inadäquat, wissenschaftliche Kausalhypothesen mit der materiellen Implikation („ $P \rightarrow Q$ “) der Logik gleichzusetzen. Man würde dann zu dem paradoxen Ergebnis kommen, daß eine Kausalhypothese notwendigerweise wahr sein muß, wenn die Prämisse (P) ganz sicher falsch ist und/oder wenn die Konklusion (Q) ganz sicher wahr ist. Etwas konkreter ausgedrückt hätte das u.a. zur Folge, daß man bei Nichtvorliegen von P keine Erwartungen hinsichtlich des Eintretens oder Nicht-Eintretens von Q ableiten kann. Dagegen ist in wissenschaftlichen Kausalhypothesen (implizit) meist auch die Aussage enthalten, daß die Folge Q nicht eintritt, wenn P nicht gegeben ist *und* wenn alle anderen Bedingungen gleichbleiben. Deshalb umfaßt die Prüfung von Kausalhypothesen immer den Vergleich von Beobachtungen unter mindestens zwei Bedingungen, die sich möglichst nur dahingehend unterscheiden, daß in einer P gegeben ist, in der anderen jedoch nicht. Wie dieses Ziel zu erreichen ist, wird das Hauptthema der folgenden Teile (insbesondere Teil 3) sein. (Näheres zum Begriff der Kausalität findet man u.a. bei Suppes, 1970; Brand, 1976; Cook & Campbell, 1979.)

Um Mißverständnisse zu vermeiden, sei betont: Eine solche Kausalhypothese bedeutet weder, daß - um im Beispiel zu bleiben - „Einstellungsänderung“ die einzige Folge des „Auftretens von Dissonanz“ sein muß, noch daß „Einstellungsänderungen“ ausschließlich durch Dissonanz hervorgerufen werden können. Aus dem meist hochkomplexen empirischen Zusammenhangsgefüge isoliert eine Kausalhypothese also in der Regel nur einen Teilaspekt.

Wir wollen uns in dieser Arbeit mit der Frage beschäftigen, wie man zu empirisch fundierten Aussagen über die „Wahrheit“ oder „Falschheit“ derartiger psychologischer Kausalhypothesen gelangen kann. Zuvor soll aber nicht versäumt werden, darauf hinzuweisen, daß Kausalaussagen der Form

---

<sup>1)</sup> Wir müssen uns in dieser Arbeit auf vereinfachende Andeutungen zur Struktur von Theorien und zur Ableitung von Hypothesen beschränken und verweisen im einzelnen auf Bunge (1967a, b), Stegmüller (1973c, 1974b, 1978, 1979a, b, 1980), Groeben & Westmeyer (1975), Suppe (1977a, b), Henning & Muthig (1979, 13-18) sowie ferner Abschnitt 2.1. Zu den Anforderungen, die eine Aussage erfüllen muß, um als wissenschaftliche Hypothese gelten zu können, siehe Bunge (1967a, 229, 280-290).

„Wenn . . . , dann . . . “ nicht die einzige Art von Hypothesen sind, die in der Psychologie interessieren. Betrachten wir als Beispiel die berüchtigte Aussage „Weiße sind (im Durchschnitt) intelligenter als Neger“. Hier wird etwas ausgesagt über die Unterschiede zwischen statistischen Parametern (Mittelwerten) zweier Populationen hinsichtlich einer bestimmten Variablen. Allgemein ausgedrückt sind solche Hypothesen Aussagen über den statistischen Zusammenhang mindestens zweier Variablen. Auf die Frage der Verursachung gehen sie nicht ein. Zur Unterscheidung von den Kausalhypothesen werden sie als *statistische Populationshypothesen* bezeichnet (Bredenkamp, 1979; Hager & Westermann, im Druck, a). Mit ihnen und anderen möglichen Arten von Hypothesen (s. Bunge, 1967a) werden wir uns gemäß unserer Hauptfragestellung nur am Rande beschäftigen.

Es gibt aber eine große Zahl möglicher *Störfaktoren*, die dazu führen können, daß eine konkrete Untersuchung oder eine bestimmte Art von Untersuchungen *nicht* als bestmögliche Prüfung einer kausalen psychologischen Hypothese bezeichnet werden kann. *Liegen in einem bestimmten Experiment derartige Störfaktoren vor, wollen wir davon sprechen, daß die Validität dieses Experiments zur Prüfung der interessierenden Kausalhypothese herabgesetzt ist.* Hieraus ergibt sich die Forderung, daß Experimente grundsätzlich so zu planen, durchzuführen und auszuwerten sind, daß ihre Validität möglichst hoch ist. Dieser Artikel befaßt sich mit den wesentlichsten Aspekten, die zu beachten sind, will man dieser Forderung annähernd nachkommen.

### 1.3 Die Validität eines Experiments

Wir werden im folgenden die angesprochenen Störfaktoren in vier Gruppen einteilen und dementsprechend vier Aspekte der experimentellen Validität behandeln:

- (1) Variablenvalidität (Teil 2)
- (2) interne Validität (Teil 3)
- (3) Situations- und Populationsvalidität (Teil 4)
- (4) statistische Validität (Teil 6)

Diese Einteilung entspricht der von Cook & Campbell (1976, 1979), die ihrerseits eine Erweiterung der Unterscheidung zwischen interner und externer Validität nach Campbell (1957, 1969) und Campbell & Stanley (1963) ist. Die folgenden Ausführungen weichen allerdings insofern grundsätzlich von diesen Ansätzen ab, als in ihnen ein Einwand berücksichtigt wird, den Gadenne (1976) gegenüber Campbell & Stanley (1963) geltend macht und der im wesentlichen auch auf die neueren Arbeiten von Cook & Campbell zutrifft: Gadenne (1976) weist nach, daß das Konzept der internen und der externen

Validität einen *induktivistischen Ansatz* darstellt, d.h. Schlüsse vom Besonderen auf das Allgemeine enthält. Da die logische Rechtfertigung induktiver Schlüsse (immer) noch aussteht, empfiehlt Gadenne (1976), das Problem der möglichen Störfaktoren im Experiment im Rahmen der Falsifikationstheorie Poppers (1976) zu behandeln, nach der zur Überprüfung von Hypothesen und Theorien ausschließlich deduktive Schlüsse anzuwenden sind.

Ohne an dieser Stelle detailliert auf wissenschaftstheoretische Probleme eingehen zu können, wollen wir im Anschluß an Gadenne (1976) in vereinfachter Weise schildern, welche Konsequenzen die Falsifikationstheorie für die Überprüfung von Kausalhypothesen wie etwa unserer WH, hat (vgl. Bredenkamp, 1980).

Das Ziel jeder empirischen Wissenschaft kann man in der Aufstellung von wahren Aussagen (Theorien, Hypothesen) mit möglichst hohem Informationsgehalt über die Realität sehen. Nach Popper (1976, 1979) kann dieses Ziel nur erreicht werden, indem man - erstens - nach Fakten sucht, die der betrachteten Theorie oder Hypothese widersprechen und - zweitens - „bessere“ Theorien aufzustellen versucht, die mit mehr empirischen Daten in Einklang stehen. Die als erstes angesprochene Überprüfung einer Theorie bzw. Hypothese erfolgt dadurch, daß man aus ihr Vorhersagen über empirisch beobachtbare Ereignisse ableitet. Diese Vorhersagen haben die Form „Wenn die Bedingung b (die sog. „Anfangsbedingung“) gegeben ist, dann tritt Ereignis e ein“. In einer experimentellen Untersuchung bezieht sich die Anfangsbedingung auf die Modalitäten der UV X, und das Ereignis e entspricht meist bestimmten Unterschieden dY auf der abhängigen Variablen. Beobachtet man nun X, ohne daß dY eintritt, ist man berechtigt, die Hypothese als „falsifiziert“ anzusehen, allerdings nur wenn der entsprechende *Basissatz* „ $X \wedge \neg dY$ “ („X hat vorgelegen, aber keine Veränderung auf Y ist eingetreten“) als reproduzierbare Tatsache akzeptiert werden kann. Praktisch bedeutet dies, daß man sich zur Falsifikation einer Hypothese nie allein aufgrund eines einzigen empirischen Ergebnisses entschließt und daß zur Prüfung einer Hypothese stets mehrere empirische Beobachtungen (konkreter: mehrere Experimente) notwendig sind. Glass (1976, 1978) und Pillemer & Light (1980) geben einen Überblick über Möglichkeiten, die Ergebnisse verschiedener Untersuchungen systematisch zusammenzufassen (s.a. Fricke, 1977; Rosenthal & Rubin, 1979; Vatz et al., 1980; Rosenthal, 1980; Cooper & Rosenthal, 1980; Bredenkamp, 1980; 35-37 und Abschn. 11.2). Zur Falsifikation wird man sich insbesondere dann entschließen, wenn eine neue Theorie aufgestellt werden kann, die auch die der alten Theorie widersprechenden Ergebnisse mit einbeziehen kann (vgl. die Beschreibung wissenschaftlicher Forschungsprogramme als Theorienketten von Lakatos (1974)). Dadurch ergibt sich ein Erkenntnisfortschritt. Solange man sich nicht für die Falsifikation einer Hypothese entschieden hat, gilt sie als „vorläufig bewährt“.

Nun sollte man von einer Bewährung der Hypothese aufgrund vorliegender Daten nur sprechen, wenn man tatsächlich versucht hat, die Hypothese auf eine „ernstzunehmende“ Weise zu widerlegen. „Ernstzunehmen“ ist ein Falsifikationsversuch, wenn er so angelegt wird, daß im Falle der Falschheit der Hypothese auch Daten zu erwarten sind, die ihr widersprechen. Es sollen also *ungerechtfertigte Bewährungen* der Hypothese vermieden werden. Ist diese Forderung erfüllt, spricht Popper (1976) von einem „*strengen*“ *Prüfversuch*. Allerdings besteht nach dieser Definition die Möglichkeit, eine strenge Prüfung dadurch zu erreichen, daß man die Untersuchung so plant, daß ausschließlich hypothesenkonträre Ergebnisse eintreten können. Dies wäre aber offensichtlich ein ungeeigneter Weg, Erkenntnisse über die Realität zu gewinnen. Deshalb soll der Verweis auf das Kriterium eines strengen Prüfversuchs immer auch folgendes bedeuten: Wenn statt der zu prüfenden Kausalhypothese WH ihr logisches Gegenteil falsch ist, soll die Wahrscheinlichkeit eines der WH widersprechenden empirischen Ergebnisses gering sein. Eine strenge Prüfung soll u.E. also sowohl fälschliche Bewährungen der WH als auch *fälschliche Falsifikationen* vermeiden.

Da diese Forderung nach strengen Prüfungen von Hypothesen und Theorien grundlegend ist für die weiteren Erörterungen in dieser Arbeit, wollen wir jetzt genauer überlegen, unter welchen Umständen eine Untersuchung zur Prüfung einer Hypothese als mehr oder weniger streng bezeichnet werden kann. Wir erinnern uns: Aus den (u.U. relativ komplexen) Anfangsbedingungen X wird mit Hilfe der zu überprüfenden WH das Ereignis dY prognostiziert. Nehmen wir nun an, es gäbe eine andere gut bewährte Hypothese HS, die aus den vorliegenden Anfangsbedingungen X (oder Teilaspekten davon) das gleiche Ereignis dY prognostiziert. Eine solche Hypothese soll als Störungshypothese bezeichnet werden. Auch wenn die WH falsch sein sollte, wäre wegen HS in diesem Fall nicht das der WH widersprechende Ereignis  $\neg dY$  zu erwarten. Deshalb kann eine Untersuchung nicht als strenger Prüfversuch bezeichnet werden, wenn in ihr die Anfangsbedingungen einer bewährten Störungshypothese vorliegen. Man kann demnach allgemein formulieren:

*Die Prüfung einer wissenschaftlichen Hypothese durch ein Experiment ist streng, wenn das aus dieser Hypothese vorhergesagte Ereignis nicht auch mit Hilfe anderer Theorien oder Hypothesen (sog. Störungshypothesen) aus Bedingungen, die im Falle dieses Experiments vorliegen, abgeleitet werden kann.*

Diejenigen Merkmale des Experiments, aus denen mit Hilfe einer Störungshypothese die gleiche Prognose abgeleitet werden kann wie aus der zu prüfenden wissenschaftlichen Hypothese, werden als *Störfaktoren* oder *Störbedingungen* der experimentellen Validität bezeichnet. *Liegt in einer Untersuchung eine potentielle Störbedingung nicht vor, heißt diese Bedingung kontrolliert.*

Eine Präzisierung des Begriffs der Strenge einer Prüfung, die es erlauben würde, jeder Untersuchung einen Zahlenwert zuzuordnen, der ein Maß dafür

ist, wie streng diese Untersuchung als Prüfung einer bestimmten Hypothese ist, liegt noch nicht vor. Wir können vielmehr nur folgendes sagen:

*Die Prüfung einer wissenschaftlichen Hypothese durch eine Untersuchung  $U_1$  ist strenger als durch eine Untersuchung  $U_2$ , wenn alle potentiellen Störfaktoren, die in  $U_1$  kontrolliert sind, auch in  $U_2$  kontrolliert sind, und wenn in  $U_1$  zusätzlich mindestens ein weiterer Störfaktor kontrolliert ist (Gadenne, 1976, 64).*

über den Begriff der Strenge eines Prüfversuchs wollen wir jetzt den Begriff der Validität eines Experiments (bzw. allgemeiner einer Untersuchung) definieren (vgl. Hager & Westermann, im Druck, a):

*Die Validität einer Untersuchung zur Überprüfung einer wissenschaftlichen Hypothese ist um so größer, je größer die Wahrscheinlichkeit ist, daß die Untersuchung Daten erbringt, die der Hypothese widersprechen, falls diese tatsächlich falsch ist, bzw. je größer die Wahrscheinlichkeit ist, daß die Untersuchung Daten erbringt, die der Hypothese nicht widersprechen, falls das logische Gegenteil dieser Hypothese falsch ist.*

Wir werden in den folgenden Abschnitten ausführlich besprechen, welche konkreten Maßnahmen ein Experimentator zu treffen hat, um die so definierte Validität seines Experiments möglichst groß werden zu lassen (vgl. dazu auch die nicht vorwiegend statistisch orientierten Ausführungen zur Planung des Experiments von Campbell & Stanley, 1963; Cochran, 1968a; Bredenkamp, 1969a, 1980, 1-40; Armitage & Remington, 1970; Stanley, 1973; Cook & Campbell, 1979; Henning & Muthig, 1979). Ganz allgemein laufen diese Maßnahmen darauf hinaus, möglichst viele der potentiellen Störungshypothesen auszuschließen, indem man die Untersuchung so plant, durchführt und gestaltet, daß die für die Anwendung dieser Störungshypothesen notwendigen Anfangsbedingungen nicht gegeben sind, daß also die entsprechenden Störfaktoren kontrolliert sind. Die dabei im folgenden zu besprechende Einteilung der potentiellen Störfaktoren in verschiedene Gruppen ist - das sei ausdrücklich betont-weder die einzig mögliche noch ist unser „Katalog“ von Störfaktoren abgeschlossen und umfassend.)

---

<sup>2)</sup> Unberücksichtigt bleiben bei der folgenden Diskussion triviale Störungen der experimentellen Validität durch falsches Verhalten des Versuchsleiters, falsches Aufzeichnen der Antworten, Fehler bei der Berechnung, bewußte Fälschung der Daten usw. (siehe Mosteller, 1968; Barber, 1976).

## 2. Variablenvalidität (VV)

In unserer exemplarischen Hypothese WH, wird eine Aussage über die Beziehung zwischen zwei Begriffen gemacht: „Dissonanz“ und „Einstellung“. Beide Begriffe sind keine empirischen Begriffe, sondern gehören zur theoretischen Sprache. In diesem Abschnitt wollen wir zunächst erläutern, was diese Einordnung bedeutet, und dann überlegen, welche Konsequenzen sie für die Validität von Experimenten zur Überprüfung von Kausalhypothesen hat.

Nach Carnap (1960) kann man die wissenschaftliche Sprache in zwei Stufen einteilen, indem man zwischen einer *Beobachtungssprache* und einer *theoretischen Sprache* unterscheidet (zur Kritik dieses Ansatzes siehe Suppe, 1977a). Die Beobachtungssprache umfaßt dabei ausschließlich Begriffe, die sich auf beobachtbare Objekte, Eigenschaften und Relationen beziehen sowie solche Begriffe, die sich durch explizite Definitionen vollständig auf diese zurückführen lassen (Stegmüller, 1974b). Wir wollen das Vokabular der Beobachtungssprache kurz als *Beobachtungsbegriffe* bezeichnen. Alle Begriffe, die nicht zur Beobachtungssprache gehören, sind *theoretische Begriffe*. Diese theoretischen Begriffe sind also nicht vollständig auf Beobachtbares zurückzuführen, sie sind nach MacCorquodale & Meehl (1948) *hypothetische Konstrukte* mit einer *Überschußbedeutung*.

Dieses Konzept ist insofern eine Idealisierung, als man inzwischen klar erkannt hat, daß keine Beobachtung voraussetzungs-, d.h. theoriefrei ist (Herrman, 1973; Lakatos, 1974; Suppe, 1977a,b). Deshalb spricht Hempel (1974) statt von einer Beobachtungssprache von einem „vorgängig verfügbaren Vokabular“.

Von besonderer Bedeutung ist nun der Umstand, daß ohne theoretische Begriffe keine Kausalaussagen möglich sind.

Zu Beginn des Abschnittes 1.2 hatten wir bereits erwähnt, daß ein Ereignis (z.B. D) dann als Ursache eines anderen Ereignisses (z.B. dE) bezeichnet werden kann, wenn beide zusammen auftreten und wenn es eine *allgemeine Gesetzmäßigkeit* gibt, wonach D (regelmäßig) dE zur Folge hat (Gadenne, 1976). Eine Aussage wie „ $D \rightarrow dE$ “ kann aber nur dann als allgemeines Gesetz bezeichnet werden, wenn es über einen bestimmten raumzeitlichen Zusammenhang hinaus gilt. Damit können die in der Aussage enthaltenen Begriffe aber nicht mehr ganz bestimmten beobachtbaren Sachverhalten entsprechen, sondern sie beziehen sich auf unbegrenzte Mengen von möglichen „Realisierungen“, die sich zumindest dadurch unterscheiden, daß sie zu unterschiedlichen Zeiten und an unterschiedlichen Orten auftreten. Die in Kausalaussagen auftretenden Begriffe müssen also notwendigerweise theoretische Begriffe sein. Zur Prüfung einer psychologischen Kausalhypothese müssen von daher stets den in ihr enthaltenen theoretischen Begriffen beobachtbare Variablen zugeordnet werden. Aussagen, die einem theoretischen Begriff einen empiri-



schen Begriff zuordnen, bezeichnet man als *Zuordnungsregeln* (Stegmüller, 1974b, 308-319).

Welche Anforderungen sind nun an diese Zuordnungsregeln zu stellen, wenn wir vom Ziel einer strengen Prüfung der Hypothese ausgehen? Oder anders ausgedrückt: Durch welche Mängel bei der Zuordnung von empirischen und theoretischen Begriffen kann die Validität einer Untersuchung als Prüfung einer Kausalhypothese eingeschränkt werden?

Wir wollen die wichtigsten dieser Mängel zu fünf Störfaktoren zusammenfassen. Da sie die Beziehung zwischen empirischen und theoretischen Variablen betreffen, wollen wir diesen Teilaspekt der Validität einer Untersuchung als „Variablenvalidität“ (W) bezeichnen und entsprechend von „Störfaktoren (VV)“ sprechen, um sie von den später noch zu besprechenden anderen Arten von Störfaktoren zu unterscheiden.

## 2.1 Mangelnde Eindeutigkeit der Zuordnung als Störfaktor (VV)

Aus einer Kausalhypothese wie  $WH_u$  ist eine empirisch prüfbare Prognose der oben beschriebenen Form  $X \wedge dY$  nur ableitbar, wenn man den theoretischen Begriffen der Hypothese ganz bestimmte empirisch beobachtbare Variablen (*empirische Realisierungen* oder *Operationalisierungen*) zuordnet. Danach muß für jede empirische Variable eindeutig entscheidbar sein, ob sie einem bestimmten theoretischen Begriff zugeordnet ist oder nicht. Liegt diese eindeutige Zuordnung nicht vor, ist es also z.B. unsicher, ob durch bestimmte Maßnahmen X „Dissonanz“ erzeugt wird und/oder ob das Ergebnis Y eines gegebenen Tests eine Entsprechung des theoretischen Begriffs „Einstellung“ darstellt, kann ein auf den ersten Blick hypotesenkonträres Untersuchungsergebnis  $X \wedge \neg dY$  darauf zurückgeführt werden, daß Y eine inadäquate Operationalisierung für die „Einstellung“ ist und/oder daß in X gar keine Entsprechung der Dissonanzbedingung D vorgelegen hat. Da in diesem Falle keine der Hypothese widersprechenden empirischen Ergebnisse auftreten können, kann von einer echten Prüfung der Hypothese gar keine Rede sein.

Die „Erklärung“ (erwartungswidriger) empirischer Ergebnisse durch Verweisen auf inadäquate Operationalisierungen findet man sehr häufig in den mit „Diskussion“ überschriebenen Teilen von Forschungsberichten. Entsprechende Argumentationen stellen nicht unbedingt eine besonders lobenswerte „kritische Würdigung“ der eigenen Forschungsbefunde dar, sondern weisen (fast) stets auf (häufig gravierende) Fehler in der Konzeption der Untersuchung hin.

Wie kann man diesen Fehler vermeiden, d.h. wie kommt man zu einer eindeutigen Zuordnung von empirischen zu theoretischen Variablen? Um es gleich vorwegzunehmen: Einen routinemäßig beschreibbaren Weg zu diesem Ziel

gibt es nicht. Wir können deshalb nur exemplarische Überlegungen skizzieren, die dem Leser als Anregungen für eigene Problemlösungen dienen mögen.

Betrachten wir eine Hypothese wie WH., isoliert, ist eine Zuordnung empirischer Variablen zu den theoretischen Begriffen recht willkürlich möglich, wenn wir einmal davon absehen, daß sie nicht unserem Vorverständnis über die verwendeten theoretischen Begriffe widersprechen sollte. Um diese Beliebigkeit einschränken zu können, benötigen wir eine *Theorie*, d.h. eine Menge von Sätzen, die die in unserer Hypothese auftretenden Begriffe dadurch näher spezifiziert, daß sie sie mit anderen theoretischen Begriffen in Beziehung setzt. Sollte durch diese Explikation noch keine eindeutige Ableitung von Operationalisierungen möglich sein, muß man ein Vorgehen wählen, das Ähnlichkeit hat mit der *Konstruktvalidierung* psychologischer Tests (Cronbach & Meehl, 1955; Campbell & Fiske, 1959):

Durch eine Menge theoretischer Sätze entsteht - um ein Bild Hempels (1974) zu benutzen - ein Netz theoretischer Begriffe, zu dem ein „passendes“ Netz empirischer Begriffe gefunden werden muß. Man ordnet empirische und theoretische Begriffe bei einer solchen Betrachtungsweise also nicht einzeln einander zu, sondern versucht, für eine Menge theoretischer Begriffe empirische Entsprechungen so zu finden, daß die nach der Theorie bestehenden Verbindungen zwischen den theoretischen Begriffen sich in den statistischen Assoziationen zwischen den entsprechenden empirischen Variablen wiederfinden. Durch einen Rekurs auf einen größeren theoretischen Zusammenhang kann also die Eindeutigkeit in der Zuordnung von empirischen zu theoretischen Begriffen erhöht werden. Allerdings sind sehr viele psychologische Theorien noch zu unpräzise formuliert, als daß man die jeweilige Menge der möglichen Operationalisierungen für ihre Begriffe tatsächlich genau spezifizieren kann. Damit sind solche Theorien und die aus ihr abgeleiteten Hypothesen aber nur in begrenztem Maße einer strengen Prüfung zugänglich.

## 2.2 Mangelnde konzeptuelle Replikation als Störfaktor (VV)

Nach der hier zugrunde gelegten Zweisprachenkonzeption von Carnap (1960) haben theoretische Begriffe gegenüber der Beobachtungssprache eine Überschußbedeutung, d.h. die Bedeutung eines theoretischen Begriffs kann durch endlich viele Beobachtungsbegriffe nicht vollständig erfaßt werden. Nun wird für einen theoretischen Begriff kaum je eine optimale empirische Entsprechung zu spezifizieren sein, in der Regel werden verschiedene mögliche empirische Realisierungen zur Auswahl stehen. So kann „Dissonanz“ in ganz unterschiedlichen praktischen Situationen auftreten, und Einstellungen werden sowohl aus verbalen Antworten wie aus dem beobachtbaren Verhalten in natürlichen Situationen ermittelt (vgl. Cook & Selltitz, 1964). Wird eine Untersuchung lediglich mit anderen Operationalisierungen wiederholt, spricht man

von einer *konzeptuellen Replikation* (Carlsmith, Ellsworth & Aronson, 1976, 64-81; Bredenkamp, 1979).

*Wir wollen die konzeptuelle Replikation als um so stärker bezeichnen, je größer die Anzahl der berücksichtigten empirischen Entsprechungen ist und je verschiedenartiger die Bereiche sind, aus denen diese Realisierungen stammen.*

Es kann durchaus der Fall auftreten, daß die zu prüfende Hypothese nur für einen Teil der möglichen empirischen Realisierungen gültig ist, für den anderen dagegen nicht. So mag unsere Hypothese  $WH_u$  bspw. zutreffen, wenn „Dissonanz“ durch einstellungskonträre Handlungen operationalisiert wird, nicht aber wenn die empirische Entsprechung einstellungskonträre Information ist. In ihrer allgemeinen Formulierung wäre die  $WH_u$  dann falsch. In einem solchen Falle ist die Wahrscheinlichkeit, bei Falschheit der Hypothese auch ein ihr widersprechendes empirisches Ergebnis zu erhalten, um so höher, je stärker die konzeptuelle Replikation ist. *Deshalb ist allgemein die Prüfung einer Hypothese um so strenger, je stärker konzeptuell repliziert wird.* Konzeptuelle Replikationen sind insbesondere dann notwendig, wenn es bereits bewährte Störungshypothesen gibt, aus denen hervorgeht, daß die geprüfte Hypothese nur für bestimmte Realisierungen gelten könnte.

Für den theoretischen Begriff, der der UV der geprüften Hypothese entspricht, kann in einer Untersuchung meist nur eine empirische Entsprechung einbezogen werden. Die konzeptuelle Replikation der AV ist zwar durch multivariate Untersuchungen mit theoretisch unbegrenzt vielen Entsprechungen der AV leichter möglich, versuchs- und auswertungstechnische Probleme setzen aber auch hier i.a. enge Grenzen (siehe Teil 8). Deshalb sind für eine strenge Prüfung einer Hypothese mehrere Untersuchungen notwendig, die sich nur hinsichtlich der vorgenommenen Operationalisierungen unterscheiden.<sup>3)</sup> Erst danach wäre eine Falsifikation der Hypothese zu rechtfertigen.

Ergeben sich bei der konzeptuellen Replikation unterschiedliche Resultate im Hinblick auf die Gültigkeit der Hypothese, braucht das nicht unbedingt zu einer generellen Verwerfung der Hypothese zu führen, sondern kann auch Anlaß sein für eine Neuabgrenzung oder Differenzierung des entsprechenden theoretischen Begriffs (Carlsmith, Ellsworth & Aronson, 1976; Bredenkamp, 1979) oder für eine Änderung der Zuordnungsregel.

Auf jeden Fall machen die bisherigen Ausführungen zur Variablenvalidität deutlich, daß kaum je eine wissenschaftliche Hypothese schon aufgrund des Ergebnisses einer einzigen empirischen Untersuchung als falsifiziert oder gut bewährt bezeichnet werden sollte.

---

<sup>3)</sup> Von den möglichen Operationalisierungen stehen dem Experimentator aus technischen oder ethischen Gründen meist nicht alle zur Verfügung (zu den ethischen Problemen beim Experimentieren siehe Klauer, 1973, 149-160; Carlsmith, Ellsworth & Aronson, 1976, 93-117; Schuler, 1980).

## 2.3 Mangelnde Entsprechung im Variationsbereich von theoretischen und empirischen Variablen als Störfaktor (VV)

Diese mögliche Beeinträchtigung der Variablenvalidität betrifft fast ausschließlich die unabhängigen Variablen. Für die ihnen entsprechenden theoretischen Variablen kann jeweils ein bestimmter Variationsbereich umschrieben werden. So kann unsere Variable „Dissonanz“ von „fehlender Dissonanz“ in - idealiter - unendlich vielen verschiedenen Abstufungen bis zu einer „extrem starken Dissonanz“ schwanken. Für eine adäquate empirische Realisierung ist ein entsprechender Variationsbereich zu fordern. Wird eine theoretische Variable im Experiment als UV operationalisiert, werden aus der großen Zahl von möglichen Ausprägungen meist nur relativ wenige berücksichtigt. Nun ist es aber durchaus möglich, daß die geprüfte Hypothese nur für bestimmte Teilmengen der Ausprägungen gültig ist (z.B. könnten nur mittlere Dissonanzen zu Einstellungsänderungen führen). Ein Experiment ist folglich um so valider, je mehr Ausprägungen der UV berücksichtigt sind bzw. je vollständiger die verwendeten Ausprägungen dem möglichen Variationsbereich der theoretischen UV entsprechen. Dabei kommt es nicht nur darauf an, daß empirische Entsprechungen für die extremen Bereiche der theoretischen Variablen vorhanden sind, vielmehr muß der dazwischenliegende Bereich repräsentiert sein, um auch U-förmige oder noch kompliziertere Beziehungen zwischen unabhängiger und abhängiger Variable entdecken zu können (vgl. Wormser, 1974).

Da die Menge der möglichen Ausprägungen der UV meist entweder unendlich oder nicht genau zu spezifizieren ist, ist die angestrebte Entsprechung nicht durch eine Zufallsauswahl der zu verwendenden Ausprägungen zu erreichen, sondern nur durch eine systematische Auswahl (z.B. je eine Behandlungsbedingung mit großer, mittlerer und fehlender Dissonanz). Dann besteht auch die Möglichkeit, neben der Hypothese über den generellen Einfluß der UV auf die AV noch spezifischere Hypothesen über die Beziehung in einzelnen Teilbereichen der UV zu prüfen. Gerade bei der Durchführung von Experimenten in der Sozialpsychologie ist es oft gar nicht so einfach, extreme Ausprägungen der theoretischen unabhängigen Variablen zu realisieren. Die im Experiment möglichen Behandlungen sprechen nämlich wegen der Künstlichkeit der Situation in der Regel die Probanden viel zu wenig an, um beispielsweise eine starke Dissonanz zu erzeugen (Carlsmith, Ellsworth & Aronson, 1976). Von daher stellen Experimente nicht unbedingt immer die strengstmögliche Prüfung einer Hypothese dar (vgl. Teil 4 und 5). Vielmehr müssen wir gerade im Experiment mit einer Störung der Variablenvalidität dadurch rechnen, daß die experimentelle Manipulation nur geringen Unterschieden auf der theoretischen UV entspricht und deshalb u.U. nicht die von der Hypothese vorhergesagte Wirkung auf die AV zeigt. In Abhängigkeit von der Art der wissenschaftlichen Hypothese (vgl. Abschn. 8.1) kann eine solche Störung der Variablenvalidität zu ungerechtfertigten Falsifikationen oder Bestätigungen führen.

## 2.4 Zu geringes Skalenniveau als Störfaktor (VV)

Wir beginnen mit einer These, die wir im folgenden erläutern und begründen:

*Die Prüfung einer wissenschaftlichen Hypothese ist um so strenger, je besser das Skalenniveau der empirischen Variablen der Struktur der theoretischen Begriffe entspricht.*

Die Struktur eines theoretischen Merkmals oder Begriffs ist festgelegt durch die Art der Relationen, die auf der Menge aller Merkmalsausprägungen definiert sind (vgl. dazu Stegmüller, 1974b). Ein Merkmal hat die einfachste Form einer Struktur, wenn man lediglich mehrere einander ausschließende und erschöpfende Ausprägungen unterscheidet, bei Einstellungen beispielsweise „konservative“, „liberale“ und „sonstige“. Geht man davon aus, daß zwischen den Ausprägungen eine Rangordnung besteht, berücksichtigt man schon ein Strukturmerkmal mehr. Man gelangt dadurch zu *komparativen Begriffen*. Die höchste uns interessierende Form der Strukturiertheit haben *metrische Begriffe*: Bei ihnen sind zwischen den einzelnen Ausprägungen auch *Abstände* definiert, man will also z.B. eine Aussage darüber machen können, ob der Unterschied in der Einstellung zu Gastarbeitern zwischen Alfred und Bruno größer ist als zwischen Claus und Dirk. Nehmen wir nun an, wir hätten dem uns interessierenden theoretischen Einstellungsbegriff als empirische Entsprechung das numerische Ergebnis  $Y$  eines genau definierten Skalierungsverfahrens zugeordnet. Bei einem komparativen Einstellungsbegriff würde das etwa der folgenden Zuordnungsregel entsprechen:

- (1) Wenn Person 1 eine positivere Einstellung zu Gastarbeitern hat als Person 2, ist  $Y_1$  größer als  $Y_2$ .

Bei einem metrischen Einstellungsbegriff muß zusätzlich noch die Entsprechung von Abständen auf der theoretischen Variablen und Differenzen auf der empirischen Variablen festgelegt werden:

- (2) Wenn der Unterschied in den Einstellungen der Person 1 und 2 größer ist als der bei den Personen 3 und 4, dann gilt  $|Y_1 - Y_2| > c |Y_3 - Y_4|$ .<sup>4)</sup>

Die Zuordnungsregel (1) ist aber nur dann sinnvoll, wenn  $Y$  eine Messung mindestens auf Ordinalskalenniveau ist (vgl. Suppes & Zinnes, 1963). Haben die  $Y$ -Werte nämlich kein Ordinal-, sondern nur Nominalskalenniveau, kön-

---

<sup>4)</sup> Zuordnungsregeln dieser Art sind keine „operationalen Definitionen“ im Sinne Bridgmans (1927), weil sie eine andere logische Struktur aufweisen. Ferner gehen wir davon aus, daß einem theoretischen Begriff durch mehrere Zuordnungsregeln mehrere empirische Variablen entsprechen können. Siehe zur Inadäquatheit von operationalen Definitionen u.a. Bunge (1967a), Herrmann (1973) und Stegmüller (1974b).

nen die von Personen zugeordneten Zahlenwerte insofern beliebig transformiert werden, als lediglich gewährleistet sein muß, daß genau die Personen, die vor der Transformation *gleiche* Zahlenwerte auf der Variablen Y hatten, auch nach der Transformation *gleiche* Zahlenwerte zugeordnet bekommen. Drückt man diese Forderung formal aus, ergibt sich: Hat die Zahlenzuordnung nur Nominalskalenniveau, sind beliebige eindeutige Transformationen erlaubt. Damit sind aber auch nicht-monotone Transformationen zulässig, und Aussagen über die Rangordnung von Zahlenwerten sind empirisch nicht sinnvoll, da ihr Wahrheitswert sich unter diesen zulässigen Transformationen ändert.

Entsprechend kann begründet werden, daß Zuordnungsregel (2) nur dann sinnvoll ist, wenn Y Intervallskalenniveau aufweist. Nur in diesem Fall bleibt nämlich bei erlaubten Transformationen der Zahlen die in der Zuordnungsregel (2) getroffene Aussage über die Größenordnung von Zahlendifferenzen in ihrem Wahrheitswert erhalten.

Welche Konsequenzen hat es für die Prüfung einer Kausalhypothese, wenn das durch die Struktur der theoretischen Begriffe und die Art der Zuordnungsregel geforderte Skalenniveau nicht gegeben ist? Nehmen wir als Beispiel an, unsere Einstellungsvariable Y habe kein Ordinalskalenniveau. Dann ist die Zuordnungsregel (1) nicht mehr sinnvoll, und die Aussage der Hypothese  $WH_u$  kann nicht mehr in die Aussage „übersetzt“ werden, daß unter der Dissonanzbedingung der Wert der Variablen Y größer ist als unter der Bedingung „keine Dissonanz“. über diese Prognose ist  $WH_u$  also nicht mehr prüfbar. Wir können diesen Sachverhalt auch anders formulieren: Wird aus einer wissenschaftlichen Hypothese eine Aussage über die Rangordnung von Einzelwerten oder Medianen auf einer empirischen Variablen abgeleitet, muß diese Variable mindestens Ordinalskalenniveau haben. Sind in der Hypothese metrische Begriffe enthalten und wird aus ihr eine Vorhersage abgeleitet, die zum Beispiel die Rangordnung arithmetischer Mittelwerte oder die Größe von Produkt-Moment-Korrelationskoeffizienten betrifft, muß mindestens Intervallskalenniveau vorliegen, da diese Aussagen ihren Wahrheitswert verändern können, wenn alle monotonen Transformationen erlaubt sind - zur Begründung siehe Suppes & Zinnes (1963) und Orth (1974).

*Allgemein gilt: Liegt das durch die Struktur der theoretischen Begriffe und die Art der aus der wissenschaftlichen Hypothese abgeleiteten Vorhersage geforderte Skalenniveau nicht vor, kann die Hypothese über diese Vorhersage nur einer weniger strengen Prüfung unterzogen werden.*

Diese Überlegungen haben wichtige Konsequenzen für die Anwendung statistischer Testverfahren. Wie im Teil 6 noch zu zeigen sein wird, werden zur Überprüfung wissenschaftlicher Kausalhypothesen aus ihnen statistische Hypothesen abgeleitet. Die üblichen parametrischen Testverfahren wie der t- und

der F-Test (s. dazu im einzelnen Teil 7) prüfen vor allem statistische Hypothesen über Gleichheit und Rangordnung von arithmetischen Mittelwerten bzw. über Produkt-Moment-Korrelationen. Diese statistischen Hypothesen stellen nach dem oben Gesagten nur sinnvolle Aussagen dar, wenn die jeweilige empirische Variable mindestens Intervallskalenniveau hat. Außerdem müssen zur mathematischen Begründung dieser Tests u.a. Annahmen über Normalverteilung und Varianzgleichheit gemacht werden, die ebenfalls nicht sinnvoll sind, wenn nur Ordinalskalenniveau vorliegt, da sich diese Verteilungsaspekte bei bestimmten monotonen Transformationen ändern (s. Abschn. 8.2). Obwohl ein bestimmtes Skalenniveau nicht zu den *mathematischen* Voraussetzungen der parametrischen Testverfahren gehört (Lord, 1953; Gaito, 1960b, 1980; Anderson, 1961; McNemar, 1962; S. 375), ist deren Anwendung und Interpretation aus den genannten Gründen nur sinnvoll, wenn mindestens Intervallskalenniveau vorliegt.<sup>5)</sup> Wie wir am Ende dieses Abschnitts sehen werden, wird dadurch ihr Anwendungsgebiet aber weniger eingeschränkt, als es auf den ersten Blick scheinen mag.

Gehen wir zunächst auf die Frage ein, wie man das Skalenniveau einer Variablen bestimmt. Nach der Definition von Suppes & Zinnes (1963) stellt eine Zuordnung von Zahlen zu Objekten dann eine Messung dar, wenn durch die Beziehungen zwischen den Zahlen empirisch beobachtbare Beziehungen zwischen den Objekten widergespiegelt werden. Sind diese empirischen Beziehungen unabhängig von jeder Zahlenzuordnung beobachtbar, kann man von einer *Repräsentationsmessung* sprechen (s. Dawes, 1977). Die hinreichenden Bedingungen dafür, daß eine solche Zahlenzuordnung Ordinal- bzw. Intervallskalenniveau hat, sind in den Axiomen sogenannter *Meßstrukturen* formuliert (Krantz et al., 1971; Orth, 1974). Um in einem konkreten Anwendungsfall das Skalenniveau zu bestimmen, muß geprüft werden, ob die empirisch beobachtbaren Beziehungen zwischen den Meßobjekten, die durch die Zahlen repräsentiert werden sollen, diese Bedingungen erfüllen (zur praktischen Durchführung dieser Prüfung siehe Westermann, 1980, im Druck, b).

Für die meisten in der Psychologie verwendeten Zahlenzuordnungen lassen sich jedoch gar keine unabhängig beobachtbaren Beziehungen zwischen Objekten auffinden, die eventuell repräsentiert werden können. Wir sprechen dann nicht von einer Repräsentations-, sondern von einer *Indexmessung* (vgl. Suppes & Zinnes, 1963; Dawes, 1977; Allerbeck, 1978). Solchen Variablen

---

<sup>5)</sup> Zur kontroversen Diskussion über die Beziehung zwischen Skalenniveau und Statistik siehe z.B. Stevens, 1951; Lord, 1953; Suppes & Zinnes, 1963; Adams, Fagot & Robinson, 1965; Baker, Hardyck & Petrinovich, 1966; Pfanzagl, 1968, S. 34-56; Gardner, 1975; Lantermann, 1976; Dawes, 1977, Kap. 7; Steinfatt, 1977; Allerbeck, 1978 sowie ferner die Reader von Heermann & Braskamp, 1970; Liebermann, 1971; Steger, 1971 und Kirk, 1972.

kann jedes gewünschte Skalenniveau zugeschrieben werden, falls sie sich nur irgendwie auf eine Abzähloperation zurückführen lassen („*Messung per fiat*“).

Postuliert man z.B. Intervallniveau für das Ergebnis eines Fragebogentests (gegeben durch die Anzahl der positiven Antworten), und ordnet man diese Variable einem metrischen Einstellungsbegriff zu, muß man sich darüber im klaren sein, daß damit die Annahme verbunden ist, daß z.B. zwischen Personen, die 4 und 6 positive Antworten gegeben haben, der gleiche Einstellungsunterschied besteht wie zwischen Personen mit 22 und 24 positiven Antworten. Man definiert damit praktisch die Abstände hinsichtlich des theoretischen Begriffs in Abhängigkeit von einer konkreten Zahlenzuordnung. Dieses Vorgehen ist nicht von vornherein unberechtigt, denn der Wissenschaftler ist ja in der Definition seiner theoretischen Begriffe relativ frei, wenn er mit Herrmann (1973) davon ausgeht, daß diese theoretischen Begriffe keine realen Entitäten mit einer zu entdeckenden Struktur bezeichnen.

Für einen Wissenschaftler, der empirisch gehaltvolle und überprüfbare Theorien aufstellen will, wird es aber immer das Ziel sein müssen, repräsentationale Messungen mindestens auf Intervallskalenniveau zu verwenden. Bei der Indexmessung drücken die Zahlen nur die unterschiedlichen Antworthäufigkeiten in einem gegebenen Erhebungsverfahren aus. Bei einer Repräsentationsmessung auf Intervallskalenniveau dagegen repräsentieren Ordnung und Abstände zwischen Zahlen empirisch beobachtbare Relationen, wie sie z.B. entstehen, wenn Probanden persönliche Präferenzen und subjektive Unterschiede hinsichtlich eines definierten Merkmals bei verschiedenen Objekten ausdrücken. Von daher ist die Zuordnung von theoretischen und empirischen Variablen bei einer Repräsentationsmessung wesentlich weniger beliebig als bei einer Indexmessung, wodurch die entsprechenden Hypothesen einer stärkeren Prüfung zugänglich werden (siehe exemplarisch für den Einstellungsbegriff Westermann, 1982).

Daß psychologische Variablen auf Intervallskalenniveau gemessen werden, ist unter anderem deshalb wichtig, weil bei Ordinalskalenniveau lediglich Aussagen über monotone Zusammenhänge zwischen Variablen gemacht werden können und Aussagen über die Art der Funktion (linear, S-förmig usw.) erst sinnvoll sind, wenn mindestens Intervallskalenniveau vorliegt.

Welche neuen Aspekte ergeben sich nun für die Prüfung statistischer Hypothesen durch parametrische Testverfahren? Zwar sollte bei subjektiven Variablen wie „Einstellung zu Gastarbeitern“, „Arbeitszufriedenheit“ oder „Angst“ die Annahme des Intervallskalenniveaus durch die empirische Prüfung meßtheoretischer Axiome fundiert werden, doch betreffen viele empirische Voraussagen, die zur Überprüfung wissenschaftlicher Hypothesen abgeleitet werden, physikalische Variablen wie „Reaktionszeit“ oder „Hautwiderstand“. Diese haben i.d. R. Verhältnisskalenniveau (zur Begründung Orth, 1974, 47-49). Andere Vorhersagen betreffen Variablen, die als einfache Abzählungen interpretiert werden können und damit zwangsläufig auf Absolutskalenni-



veau liegen. So könnte - als Beispiel zum letzten Punkt - aus einer Kausalhypothese abgeleitet werden, daß die Anzahl der positiven Antworten zu Gastarbeitern im Fragebogen FB um so größer ist, je stärker die erzeugte Dissonanz war. Falls keine zu starken Abweichungen von den mathematischen Voraussetzungen für die parametrischen Tests vorliegen (s. Abschn. 8.2), können sie auch in diesem Fall verwendet werden, um die abgeleitete statistische Hypothese zu überprüfen. Überhaupt kein Hindernis für die Anwendung parametrischer Tests stellen Probleme des Skalenniveaus dar, wenn man keine wissenschaftliche Hypothese prüft, die in theoretischen Begriffen formuliert ist, sondern vielmehr die statistische Hypothese selbst als wissenschaftliche Hypothese betrachtet (s. Abschn. 1.2).

Dies ist der Fall, wenn man z.B. nicht prüfen will, ob bei „Dissonanz“ die „Einstellungen“ positiver werden, sondern nur wissen will, ob unter einer bestimmten Bedingung A die Werte im Einstellungstest ET höher sind als unter der Bedingung nicht-A. Da hier nur der Zusammenhang von ganz bestimmten, genau definierbaren empirischen Variablen untersucht wird, sind bezüglich der AV keinerlei Transformationen sinnvoll, und man kann diese Variable zur Intervall- oder Verhältnisskala deklarieren (Messung per fiat).

## 2.5 Konfundierung von theoretischen Begriffen als Störfaktor (VV)

Bisher haben wir für eine strenge Prüfung u.a. gefordert, daß die Menge der möglichen empirischen Realisierungen für einen theoretischen Begriff genau abzugrenzen ist. Das schließt nicht aus, daß eine bestimmte empirische Variable als Operationalisierung für mehrere theoretische Begriffe betrachtet werden kann. Man bezeichnet diese Begriffe dann als *konfundiert*. In diesem Fall ist es leichter möglich, gut bewährte Störungshypothesen zu finden, die die gleiche Prognose erlauben wie die zu prüfende Hypothese. Dadurch kann die Möglichkeit einer strengen Prüfung dieser Hypothese stark gefährdet werden. Wir wollen dies anhand von vier Fallgruppen näher untersuchen.

1. Recht häufig wird in der Psychologie für eine gegebene Menge experimenteller Bedingungen, die Operationalisierungen unterschiedlicher Ausprägungen eines theoretischen Begriffs darstellen sollen, eine *alternative theoretische Interpretation* gegeben.

So führt beispielsweise Bern (1972, 1975) die Einstellungsänderungen in den Experimenten zur Prüfung von Festingers Dissonanztheorie nicht auf Unterschiede in der erzeugten Dissonanz zurück, sondern geht davon aus, daß experimentelle Bedingungen, die nach Festinger unterschiedlichen Ausmaßen an „kognitiver Dissonanz“ entsprechen, zumeist für die Probanden unterschiedliche „Selbstwahrnehmungen“ mit sich bringen, die dann zu unterschiedlichen „Einstellungen“ führen.

Eine solche Konfundierung theoretischer Begriffe schränkt aber die Möglichkeit strenger Prüfungen aller beteiligten Hypothesen und Theorien nicht ein, vielmehr wird durch sie überhaupt erst die Möglichkeit eröffnet, bei empirischer Nichtbewährung einer Hypothese oder Theorie  $T_1$  diese empirischen Ergebnisse im Rahmen einer anderen Theorie  $T_2$  mit anderen theoretischen Begriffen zu erklären. Diese Art von Konfundierung ist also kein Mangel, sondern Voraussetzung für Erkenntnisfortschritt (vgl. Abschn. 1.3).

2. Auch die abhängige Variable einer Untersuchung entspricht kaum je eindeutig dem theoretischen Begriff, dem sie zugeordnet ist. In unterschiedlichen Werten der Personen auf der AV können sich vielmehr nicht nur unterschiedliche Ausprägungen der entsprechenden theoretischen Variablen ausdrücken, sondern auch Unterschiede auf einer ganzen Reihe von anderen theoretischen Variablen. So können nicht nur positivere Ausprägungen der Variablen „Einstellung zu Gastarbeitern“ zu höheren Werten im entsprechenden Einstellungstest führen, sondern beispielsweise auch stärkere Ausprägungen von Variablen wie „Antworten gemäß sozialer Erwünschtheit“, „Ja-Sage-Tendenz“ und „Tendenz zu extremen Antworten“. Daß solche sog. *Reaktionsstile* („response sets“) mit der in der Untersuchung interessierenden theoretischen abhängigen Variablen konfundiert sind, ist typisch für die in der Sozialforschung verbreiteten verbalen Skalierungsmethoden. Man kann den störenden Einfluß dieser Art von Konfundierung auf die Validität der Untersuchung verringern, indem man im Rahmen einer konzeptuellen Replikation (siehe Abschn. 2.2) auch „nicht-verbale“ Realisierungen des entsprechenden theoretischen Begriffs einbezieht und/oder indem man das Erfassungsinstrument so konstruiert, daß diese Antworttendenzen sich möglichst wenig auswirken können (siehe dazu Edwards, 1957a,b, 1970; Cook & Selltitz, 1964; Berg, 1967; Scott, 1968; Holm, 1975, 1976, 1977; Koch, 1976).

Auch wenn die *Reliabilität* der empirischen Variablen gering ist, liegt eine Störung der Variablenvalidität durch Konfundierung theoretischer Begriffe vor: Im Sinne der klassischen Testtheorie (vgl. Lord & Novick, 1968; Fischer, 1974; Kranz, 1979; Wottawa, 1980) sind dann in den beobachteten Werten der empirischen Variablen Meßfehler mit den sog. „wahren“ Werten in einem zu großen Maße vermengt.

3. Nach einer hauptsächlich auf Orne (1962) zurückgehenden Auffassung verhalten sich Teilnehmer eines Experimentes gemäß ihren Hypothesen darüber, was im Experiment von ihnen erwartet wird, was Ziel und Sinn dieses Experimentes ist.<sup>6)</sup> Die Art der von den Versuchspersonen (Vpn) gebildeten Hypo-

---

<sup>6)</sup> Zu Einzelheiten über Forschungen zur Sozialpsychologie des Experimentes siehe z.B. Rosenthal & Rosnow, 1969a; Klauer, 1973; Timaeus, 1974, 1975; Kruglanski, 1975; Mertens, 1975; Barber, 1976; Gniech, 1976; Rosnow & Davis, 1977; Silverman, 1977; Rosenthal & Rubin, 1978; Bungard, 1980; Callaway, Nowicki & Duke, 1980.

thesen hängen von gewissen „*demand characteristics*“ der experimentellen Situation ab. Eine dieser Determinanten der Probanden-Hypothesen sind die Erwartungen des Versuchsleiters, deren Auswirkungen ausführlich von Rosenthal (1969, 1977) untersucht wurden. Werden nun die Ausprägungen einer theoretischen Variablen durch unterschiedliche experimentelle Bedingungen operationalisiert, haben diese Bedingungen unterschiedliche „*demand characteristics*“ und können folglich zu unterschiedlichen Hypothesen bei den Vpn führen. Unter der Gültigkeit der Auffassung von Orne wären also die theoretische Variable, die der experimentellen UV entspricht, und die Variable „Hypothesen der Vpn“ zwangsläufig konfundiert. Nun ist zwar die Annahme, daß das Verhalten im Experiment generell von den Hypothesen der Vpn bestimmt wird, empirisch kaum zu stützen (Bredenkamp 1980, 41-47), doch muß man in einzelnen Untersuchungen immer mit solchen Effekten rechnen. Um eine möglichst strenge Prüfung einer Hypothese zu erreichen, muß man sich folglich bemühen, die Bildung entsprechender VP-Hypothesen zu erschweren, bzw. man muß verhindern, daß eine eventuelle Hypothesenbildung bei den Vpn die gleiche empirische Auswirkung hat, die auch von der zu prüfenden Kausalaussage vorhergesagt wird. In unserem Fall muß etwa verhindert werden, daß die Vpn unter der Dissonanz-Bedingung leicht erkennen können, daß von ihnen eine Einstellungsänderung erwartet wird.

Es gibt eine Reihe von Techniken, die verhindern sollen, daß Probanden unter verschiedenen experimentellen Bedingungen unterschiedliche Hypothesen über das von ihnen erwartete Verhalten bilden (siehe Carlsmith, Ellsworth & Aronson, 1976, 280-301 und Rosnow & Davis, 1977):

- man vermeidet, daß Personen Unterschiede zwischen den Behandlungsbedingungen überhaupt wahrnehmen können (Blindversuch, ähnlich den Placebo-Experimenten in der Pharmaforschung),
- man verleitet alle Personen dazu, falsche Hypothesen zu bilden,
- man erhebt die AV in einer anderen Situation,
- man täuscht die Personen über ihre Rolle als Versuchspersonen, z.B. indem man sie glauben macht, sie seien die Versuchsleiter,
- man führt ein Feldexperiment (s. Abschn. 3.6) durch, in dem die Personen nicht bemerken, daß sie Teilnehmer an einer wissenschaftlichen Untersuchung sind,
- man untersucht als AV statt leicht zu gebender und folgenloser verbaler Antworten offene Verhaltensweisen, die für die Probanden auch subjektiv bedeutsam sind,
- man vermeidet wiederholte Erfassungen der AV an den gleichen Personen (s.a. Abschn. 3.2).
- man prüft Hypothesen, die schwierig zu „erraten“ sind,
- man bittet die Probanden um ehrliches und unvoreingenommenes Antwortverhalten.

Vom Experimentator ausgehende störende Beeinflussungen des Versuchspersonen-Verhaltens können vor allem dadurch eingeschränkt werden, daß der Versuchsleiter die geprüfte Hypothese nicht kennt (naiver Versuchsleiter) und/oder daß er nicht weiß, welcher experimentellen Bedingung die jeweils untersuchte Person zugeordnet ist (Doppelblindversuch). Auch eine Standardisierung der Kommunikation zwischen Versuchsleiter und -personen kann die Gefahr einer solchen Validitätsstörung vermindern.

4. Wir wollen jetzt eine Art von Konfundierung ansprechen, deren Ausmaß davon abhängt, wie der Experimentator die experimentellen Modalitäten definiert. Stellen wir uns dazu zwei unterschiedliche empirische Realisierungen von „starker Dissonanz“ ( $X_1$ ) und „fehlender Dissonanz“ ( $X_2$ ) für eine Person mit einer negativen Einstellung zu Gastarbeitern vor. (Im folgenden beziehen sich die eingeklammerten Indices auf die beiden Fälle 1 und 2.)

- Fall 1:  $X_{(1)1}$ : Einen Abend mit einer sehr angenehmen Gastarbeiterfamilie verbringen.  
 $X_{(1)2}$ : Einen negativen Text über Gastarbeiter lesen.
- Fall 2:  $X_{(2)1}$ : Einen sehr positiven Text über Gastarbeiter lesen.  
 $X_{(2)2}$ : Einen negativen Text über Gastarbeiter lesen.

Die Bedingungen  $X_{(1)1}$  und  $X_{(1)2}$  unterscheiden sich nicht nur hinsichtlich der „Dissonanz“, sondern auch noch hinsichtlich der empirischen Entsprechungen einer großen Zahl anderer theoretischer Variablen. Demzufolge sind im Fall 1 eine Fülle von Störungshypothesen denkbar, die die gleiche Prognose erlauben wie die  $WH_u$ . So könnte eine positivere Einstellung unter  $X_{(1)1}$  als unter  $X_{(1)2}$  durch die recht gut bewährte Hypothese vorhergesagt werden, daß Einstellungsänderungen durch persönliche Kontakte eher erfolgen als durch verbale Information. Auch im Fall 2 können sich die Bedingungen noch hinsichtlich anderer Variablen als „Dissonanz“ unterscheiden (z. B. „Verständlichkeit“ oder „Glaubwürdigkeit“ etc.), ihre mögliche Anzahl ist aber doch weit geringer als im Fall 1. Daher sind auch weniger bewährte Störungshypothesen zu erwarten. Im Fall 2 liegt also eine strengere Prüfung unserer Hypothese  $WH$ , vor als im Fall 1.

In diesem Beispiel wird deutlich, daß zur strengen Prüfung einer Hypothese die experimentellen Bedingungen, die empirische Realisierungen von verschiedenen Ausprägungen der theoretischen unabhängigen Variablen darstellen, sich so weit ähneln sollen, daß hinsichtlich möglichst weniger anderer Variablen systematische Unterschiede bestehen. Dieses Ziel ist allerdings besonders in sozialpsychologischen Untersuchungen mit relativ komplexen theoretischen Variablen recht schwer zu erreichen (siehe Carlsmith, Ellsworth & Aronson, 1976, 61-64, sowie ferner Underwood & Shaughnessy, 1975, 28-36).

## 2.6 Zusammenfassung

Wir haben in diesem Abschnitt eine erste Gruppe von Faktoren kennengelernt, die dazu führen können, daß eine Untersuchung keine strenge Prüfung der gegebenen Hypothese darstellt. Da diese Faktoren die Beziehung zwischen den in der Untersuchung auftretenden beobachtbaren Begriffen (Variablen, Bedingungen) und den entsprechenden theoretischen Variablen aus der zu prüfenden Hypothese betreffen, haben wir sie als Störfaktoren der Variablen-Validität - Störfaktoren (VV) - bezeichnet.

Nach unseren Überlegungen ist eine Kausalhypothese um so strenger prüfbar,

- je eindeutiger den theoretischen Begriffen der Hypothese empirische Variablen („Operationalisierungen“) zugeordnet sind,
- je mehr Operationalisierungen der theoretischen Begriffe berücksichtigt werden und je unterschiedlicher diese Operationalisierungen sind,
- je mehr die berücksichtigten Ausprägungen der UV den möglichen Ausprägungen der entsprechenden theoretischen Variablen entsprechen,
- je eher das Skalenniveau der empirischen Variablen der Struktur der zugehörigen theoretischen Begriffe entspricht,
- je weniger bei den berücksichtigten Operationalisierungen andere theoretische Variablen mit den Begriffen der Hypothese konfundiert sind.

(Die Konstanz aller anderen Bedingungen ist jeweils vorausgesetzt.)

## 3. Interne Validität

Wie wir im Abschnitt 1.2 gesehen haben, folgt aus der Bedeutung von Kausalaussagen, daß zu ihrer Überprüfung Beobachtungen unter mindestens zwei Bedingungen nötig sind, die sich möglichst nur dahingehend unterscheiden, daß in einer Bedingung eine Entsprechung der in der Kausalaussage spezifizierten Ursache vorliegt, in der anderen jedoch nicht. Wir wollen das am Beispiel unserer Hypothese  $WH_u$  kurz erläutern.

Um diese  $WH_u$  prüfen zu können, kann im Rahmen eines einfachen Versuchsplanes (Vpl.) oder Designs wie folgt verfahren werden: Zunächst wird bei der Person die AV ohne Dissonanz ( $X_1$ ) gemessen (Variable  $Y_{11}$ ), anschließend wird Dissonanz erzeugt ( $X_2$ ) und daraufhin wird die AV erneut gemessen ( $Y_{12}$ ). Bei einem derartigen Vorgehen spricht man von „intraindividuelle Bedingungsvariation“, die folgendermaßen symbolisiert werden kann:

(Vpl.)    1)     $X_1$      $Y_{11}$      $X_2$      $Y_{12}$

Wählt man dagegen die sog. „interindividuelle Bedingungsvariation“, ist wie folgt vorzugehen: Bei einer Person wird Dissonanz erzeugt ( $X_2$ ), bei einer

anderen jedoch nicht ( $X_1$ ), und die Einstellung der Personen wird gemessen ( $Y_{22}$  und  $Y_{11}$ ). Der resultierende Versuchsplan ergibt sich wie folgt:

(Vpl.2)	2)	$X_1$	$Y_{11}$
		$X_2$	$Y_{22}$

Unter Berücksichtigung der Zuordnungsregel (1) aus Abschnitt 2.4 wird bei Gültigkeit der Hypothese  $WH_u$  erwartet, daß  $Y_{12}$  größer als  $Y_{11}$  ist. Die gleiche Vorhersage läßt sich jedoch auch mit Hilfe anderer Hypothesen ableiten, falls sich die Messungen  $Y_{11}$  und  $Y_{12}$  nicht nur dadurch unterscheiden, daß vor  $Y_{12}$  die Dissonanzbedingung realisiert wurde, vor  $Y_{11}$  dagegen nicht.

Wir wollen diese Bedingungen, die ebenfalls zu unterschiedlichen Werten auf der AV Y führen können, als „*Störfaktoren der internen Validität*“ bezeichnen.<sup>7)</sup> Sie können zusätzlich zu den experimentellen Bedingungen ( $X_1$  versus  $X_2$ ) wirksam werden oder aber an ihrer Stelle. Die wichtigsten dieser Störfaktoren haben wir zu zwei Gruppen zusammengefaßt, auf die wir im folgenden eingehen - eine etwas andere Einteilung findet sich bei Campbell & Stanley (1963) und bei Cook & Campbell (1979).

### 3.1 Variation personaler und situationaler Merkmale als Störfaktoren (IV)

#### 3.1.1 Variation situationaler Merkmale

Beim Vpl. 1 kann zwischen erster und zweiter Messung neben  $X_2$  noch ein anderes Ereignis aufgetreten sein, das nach einer bewährten Störungshypothese die gleiche Auswirkung hat wie nach der zu prüfenden Hypothese die experimentelle Behandlung  $X_2$ .

Bspw. könnte der Proband während der Untersuchung zur Prüfung unserer Hypothese  $WH_u$  sehr attraktive Gastarbeiterinnen kennengelernt haben, woraus nach der bewährten Störungshypothese (vgl. Amir, 1969) „angenehme persönliche Begegnungen mit Menschen aus einer bestimmten Gruppe verändern die Einstellung zur entsprechenden Gruppe zum Positiven hin“ die gleiche Einstellungsänderung folgt wie aus der  $WH_u$ ,

Entsprechend können beim Vpl. 2 sich die Situationen, in denen die beiden Messungen  $Y_{11}$  und  $Y_{22}$  erhoben wurden, über den Unterschied zwischen  $X_1$  und  $X_2$  hinaus noch hinsichtlich weiterer Merkmale unterscheiden. Liegt für

<sup>7)</sup> Wir können diese Bezeichnung von Campbell & Stanley (1963) übernehmen, weil dieser Aspekt der experimentellen Validität auch nach unserer Betrachtungsweise insofern „intern“ ist, als (wie wir noch sehen werden) zur Kontrolle dieser Störfaktoren nur das gerade durchgeführte Experiment betrachtet werden muß, während z.B. eine hohe Variablenvalidität in aller Regel mehrere Untersuchungen erfordert.

mindestens eines dieser Merkmale eine bewährte Störungshypothese vor, die die gleiche Prognose erlaubt wie die  $WH_u$ , ist die interne Validität der Untersuchung nicht gewährleistet.

### 3.1.2 Variation personaler Merkmale

Zwischen den beiden Messungen beim Vpl. 1 kann es unabhängig vom Eintreten von  $X_2$  zur Veränderung eines Merkmals des Probanden kommen, aus der sich die Einstellungsänderung vorhersagen läßt. Bspw. kann das allgemeine Aggressivitätsniveau des Probanden sinken.

Beim Vpl. 2 ist die interne Validität gestört, wenn sich die Probanden unter den beiden experimentellen Bedingungen hinsichtlich eines Merkmals unterscheiden und wenn dieser Unterschied nach einer bewährten Hypothese das auch aus der  $WH_u$  vorhergesagte empirische Ereignis zur Folge hat.

Beispiele für mögliche Störungshypothesen sind etwa: In der Bedingung  $X_2$  liegt schon vor der Untersuchung eine positivere Einstellung vor; der Proband unter  $X_2$  tendiert stärker dazu, sozial erwünschte Antworten zu geben.

## 3.2 Störfaktoren (IV) bei Meßwiederholung

Neben den bereits erwähnten, bei allen Designs möglichen Störfaktoren (IV), die in Unterschieden zwischen  $X_1$  und  $X_2$  hinsichtlich situationaler und personaler Merkmale bestehen, sind bei Designs mit wiederholter Messung der AV unter den verschiedenen experimentellen Bedingungen an den gleichen Personen (intraindividuelle Bedingungsvariation, s. Vpl. 1) weitere Gruppen von möglichen Störfaktoren (IV) zu beachten (nach Campbell & Stanley, 1963; Namboodiri, 1972; Greenwald, 1976; Cook & Campbell, 1979; s.a. Edwards, 1971, 225-226; Keppel, 1973, 395-400).

### (1) Veränderungen beim Meßinstrument

Bei der zweiten Messung der AV kann sich das Meßinstrument verändert haben. Zum Beispiel können sich bei Erfassung der AV durch menschliche Beurteiler (Rater) deren Beurteilungskriterien mit der Zeit (oder zunehmender Beurteilungspraxis) verschieben.

### (2) Ausscheiden von Versuchspersonen

Besonders wenn die Untersuchung sich über einen längeren Zeitraum hinzieht, können zwischen erster und zweiter Messung einige Versuchspersonen ausscheiden.

## (3) Sensitivierung

Vpn können durch die erste Behandlung  $X_1$  so sensitiviert werden, daß sie auf  $X_2$  anders reagieren, als wenn sie allein  $X_2$  ausgesetzt worden wären.

## (4) übungs-, Ermüdungs- und Erinnerungseffekte

Bei der zweiten Messung kann der Wert der AV bspw. durch die größere Vertrautheit mit den Aufgaben, durch eine von der ersten Messung herrührende Ermüdung oder auch durch die Erinnerung an die Antworten bei der ersten Messung mitbestimmt werden.

## (5) Vermengung von Behandlungswirkungen

Der Wert der AV kann bei der zweiten Messung außer durch  $X_2$  auch durch die zurückliegende experimentelle Behandlung  $X_1$  beeinflußt werden.

Die beiden zuletzt angesprochenen Störfaktoren (IV) werden oft als „Carry-over-“ oder „Übertragungseffekte“ zusammengefaßt. Zur Erläuterung und Unterscheidung diene ein Beispiel von Greenwald (1976, 318): Zur Untersuchung der Wirkung zweier Drogen auf die Reaktionszeit verabreicht ein Experimentator (E) den Vpn nacheinander beide Drogen und läßt anschließend jeweils einen Reaktionstest durchführen. Übungseffekte können dabei auftreten, wenn die im Test geforderten motorischen Tätigkeiten beim zweiten Mal besser beherrscht werden als beim ersten Mal. Eine Vermengung von Behandlungswirkungen liegt dagegen vor, wenn die Wirkung der zuerst verabreichten Droge bei der zweiten Einnahme noch nicht vollständig abgeklungen ist.

Zusammenfassend wollen wir die 5 Störfaktoren (IV) bei Meßwiederholung als *Sequenzeffekte* bezeichnen. Liegt ein Sequenzeffekt vor, kann über eine gut bewährte Störungshypothese die gleiche Prognose abgeleitet werden wie aus der zu prüfenden Kausalhypothese. In diesem Fall ist die interne Validität der Untersuchung gestört.

Diese Störfaktoren gelten nicht nur für den bisher betrachteten Fall einer zweimaligen Messung, sondern auch wenn eine Person  $s$  unter  $K$  experimentellen Bedingungen beobachtet wird:

(VPl. 3)       $X_1 \quad Y_{s1} \quad X_2 \quad Y_{s2} \quad \dots \quad X_K \quad Y_{sK}$

Hier entstehen i.a. zudem Sequenzeffekte höherer Ordnung, d.h. Beeinflussungen von weiter auseinanderliegenden Behandlungen und Messungen.

Ein weiterer und möglicherweise gravierender Nachteil von Designs mit wiederholten Messungen soll hier ebenfalls erwähnt werden, obwohl er eher zu den Störungen der Variablenvalidität zu rechnen ist: Eine wiederholte Messung, d.h. eine Beobachtung einer Person unter allen experimentellen Bedingungen, erleichtert die Bildung von Hypothesen über das Ziel des Experi-



ments, und es besteht daher eine erhöhte Gefahr, daß die unabhängige Variable der zu prüfenden Kausalhypothese nur konfundiert mit diesen Vermutungen des Probanden realisiert werden kann (siehe auch Abschnitt 2.5).

Weshalb werden in der Psychologie trotz dieser zusätzlichen Gefahren für die Validität der Untersuchung wiederholte Messungen der Personen unter allen experimentellen Bedingungen durchgeführt?

Zum einen bieten sich Meßwiederholungen an, wenn die hier als störend klassifizierten Übungs- und Übertragungseffekte selbst Gegenstand der Hypothesenprüfung sind, wie dies etwa in Untersuchungen zum Lernfortschritt oder zu Adaptationsphänomenen der Fall ist (vgl. Greenwald, 1976, 316-318). Zum anderen werden Designs mit wiederholten Messungen wegen ihrer Ökonomie angewendet: Eine Versuchsperson „liefert“ mehr Daten, und die für Instruktion, Übungsphase usw. notwendige Zeit ist geringer, als wenn jedes Beobachtungsdatum an einer anderen Versuchsperson erhoben werden würde.

Auf weitere Vor-, aber auch Nachteile werden wir ausführlicher im Abschnitt 8.4.6 zu sprechen kommen.

### 3.3 Zur Kontrolle der Störfaktoren (IV) bei interindividueller Bedingungsvariation

Glücklicherweise brauchen wir bei der Durchführung von Untersuchungen die jeweilige Untersuchungssituation nicht nach den unendlich vielen potentiellen Störfaktoren (IV) abzuklopfen, um die möglichen Störungshypothesen einzeln auszuschalten. Vielmehr kann man diese Störfaktoren schon durch eine geeignete Planung der Untersuchung kontrollieren: durch Konstanthaltung bzw. Elimination und durch die Zufallsordnung von Probanden zu den einzelnen Behandlungsbedingungen (Randomisierung).

Wir können in diesem Artikel nur allgemein aufzeigen, durch welche Arten von Maßnahmen die Störfaktoren (IV) zu kontrollieren sind, und dies durch Beispiele illustrieren. Konkrete Handlungsanweisungen in bezug auf die Gestaltung der Untersuchungssituation (von der Auswahl der Stimuli über ihre Darbietungsform und die physikalische Ausschaltung anderer Variablen bis hin zu den Verhaltensvorschriften für den Versuchsleiter) können aus diesen allgemeinen Prinzipien für jedes Einzelexperiment abgeleitet werden. Diesbezügliche Hinweise bieten darüber hinaus die eher praktisch orientierten Einführungen in die experimentelle Psychologie (insbesondere Zimny, 1961; Selg, 1975; Carlsmith, Ellsworth & Aronson, 1976; außerdem z.B. Bugelski, 1960; Underwood, 1966; Heckhausen, 1969; Matheson, Bruce & Beauchamp, 1970; Sheridan, 1971; Arnold, 1972; Runkel & McGrath, 1972; Traxel, 1974; Wormser, 1974; Massaro, 1975; Fromkin & Streufert, 1976; Preiser, 1977; McGuigan, 1979, sowie Kazdin, 1980).

### 3.3.1 Konstanthaltung und Elimination

Wird im konkreten Fall vermutet, daß ein bestimmter Faktor zu einer Beeinträchtigung der internen Validität führt, so kann man diese mögliche Störung dadurch ausschalten, daß man den betreffenden Faktor konstant hält, und zwar sowohl *während* der gesamten Untersuchungszeit als auch - bei interindividueller Bedingungsvariation - *zwischen* den Behandlungsbedingungen (Zimny, 1961; Bredenkamp, 1969a).

Die Konstanthaltung als Technik zur Erhöhung der internen Validität ist besonders nutzbringend bei Faktoren einsetzbar, die zu den vom Experimentator hergestellten situativen Bedingungen gehören und die nicht die unabhängige Variable im engsten Sinne darstellen, also etwa bei der Instruktion, der Darbietungszeit und -form, dem Lärmpegel, dem Versuchsleiter usw. Variieren diese Faktoren systematisch zusammen mit den Modalitäten der UV, wird die interne Validität beeinträchtigt, weil die Variation der Werte auf der AV nicht mehr *ausschließlich* auf die *systematische Variation der UV* (und ggf. einen unsystematischen Fehleranteil) zurückgeführt werden kann. Man kann im übrigen diese Rahmenbedingungen durch Automatisierung des Versuches ablaufes recht problemlos konstant halten, also etwa durch elektronische Zeitsteuerung, Instruktion vom Tonband etc.

Die Elimination einer potentiellen Störbedingung kann als Spezialfall der Konstanthaltung angesehen werden. Sie ist etwa dann indiziert, wenn eine Konstanthaltung (etwa des Ausmaßes und der Qualität der sozialen Kontakte) nicht erreichbar ist.

Wird eine Untersuchung mit Personen in deren natürlicher Umgebung durchgeführt, gibt es eine Fülle von möglichen Störfaktoren. Ihre Konstanthaltung oder Elimination ist in den meisten Fällen unmöglich. Damit wäre aber auch eine strenge Prüfung der Hypothese nicht möglich. Von daher empfiehlt es sich, Untersuchungen in streng kontrollierten Situationen durchzuführen, also in einem gegen möglichst viele potentielle Störfaktoren abgeschirmten „Laboratorium“.

(Daß das *ausschließliche* Experimentieren im psychologischen Labor die strenge Prüfung einer Hypothese verhindern kann, wird später in Zusammenhang mit der Situationsvalidität erörtert.)

Was dabei unter Abschirmung konkret zu verstehen ist, hängt vom Bereich ab, aus dem die untersuchte Hypothese stammt. So sind in wahrnehmungspsychologischen Experimenten in erster Linie äußere Sinnesreize (plötzliche Geräusche, Beleuchtungsschwankungen usw.) konstantzuhalten oder zu eliminieren (z. B. durch einen schallisolierten Raum), in sozialpsychologischen Experimenten müssen eher die sozialen Interaktionen mit anderen Versuchsbeteiligten gleichförmig gestaltet werden und Kontakte mit Außenstehenden verhindert werden.

### 3.3.2 Randomisierung

Durch Konstanthaltung oder Elimination kann man nur solche potentiellen Störfaktoren kontrollieren, die bekannt und beobachtbar sind oder die durch so grobe Isolierungsmaßnahmen wie die Durchführung einer Untersuchung im Laboratorium ausgeschaltet werden können. Außerdem ist die Konstanthaltung oder Elimination eines Störfaktors relativ aufwendig, so daß solche Maßnahmen nur für eine relativ geringe Anzahl potentieller Störfaktoren durchführbar sind. Nun können sich die Personen, die bei einer interindividuellen Bedingungsvariation unter den verschiedenen Ausprägungen der UV untersucht werden, aber hinsichtlich einer großen und unbekannten Zahl individueller Merkmale unterscheiden.

Es gibt nur einen Weg, diese möglichen Störfaktoren (IV) zu kontrollieren: Zu jeder Modalität der UV werden *mehrere* „Untersuchungseinheiten“ (kurz: „Vpn“ für „Versuchspersonen“) beobachtet, und Vpn und Modalitäten werden *zufällig* einander zugeordnet. Dieses Vorgehen bezeichnet man als *Randomisierung*.

Nur wenn eine Randomisierung durchgeführt wird, können Kausalhypothesen streng geprüft werden: Nach einer Zufallszuordnung von Vpn und Bedingungen gibt es keine plausible Begründung für Störungshypothesen, nach denen sich die unter den verschiedenen Bedingungen beobachteten Vpn hinsichtlich eines ihrer Merkmale systematisch unterscheiden. Die Wahrscheinlichkeit, daß auf einer mit den Vpn zusammenhängenden Variablen zwischen den Behandlungsgruppen ein Unterschied von einer bestimmten Mindestgröße auftritt, ist nämlich um so geringer, je größer die Anzahl  $n$  der jedem Treatment zufällig zugeordneten Vpn ist. Zwischen den Behandlungsgruppen sind also auch nach einer Randomisierung systematische Unterschiede möglich; sie sind jedoch um so unwahrscheinlicher, je mehr Vpn verwendet werden.<sup>8)</sup>

Der Versuchsplan 2 wird mit Randomisierung zum weitverbreiteten Versuchsplan mit einer Experimental- und einer Kontrollgruppe, der i.a. als *Zufallsgruppenversuchsplan* bezeichnet wird:

(Vpl.	4)	R	$X_1$	$Y_{i1}$	
		R	$X_2$	$Y_{i2}$	(„R“ bedeutet „Randomisierung“)

Diesen Versuchsplan kann man auch in der folgenden Weise darstellen, die sich für die weiteren Erörterungen als nützlich erweisen wird (Vpl. 5):

---

<sup>8)</sup> Diese zufällig möglichen Unterschiede werden bei der Versuchsauswertung über Signifikanztests berücksichtigt (vgl. Teil 7). Wie die nach bestimmten Kriterien optimale Anzahl der Vpn zu bestimmen ist, stellen wir im Teil 10 dar.

(Vpl. 5)

R	– UV X
X <sub>1</sub>	X <sub>2</sub>
Y <sub>11</sub>	Y <sub>12</sub>
Y <sub>21</sub>	Y <sub>22</sub>
⋮	⋮
Y <sub>n<sub>1</sub>1</sub>	Y <sub>n<sub>2</sub>2</sub>

Dabei bezeichnen die  $Y_{ik}$ -Werte die Größe der AV bei den verschiedenen Vpn. Aus dem Vpl. 5 wird ersichtlich, daß in der sog. „Kontrollgruppe“ ( $X_1$ ) die Anzahl der Vpn gleich  $n_1$  und in der „Experimentalgruppe“ ( $X_2$ ) gleich  $n_2$  ist, so daß also insgesamt  $n_1 + n_2 = N$  Vpn untersucht werden.

Bei multifaktoriellen Experimenten werden die Vpn nach dem Prinzip der Randomisierung zufällig den möglichen *Kombinationen* von Modalitäten zugewiesen.

Beispiel: In einem Experiment zur Prüfung unserer  $WH_{II}$  könnte eine UV B vier Dissonanzstärken entsprechen (z.B. „fehlende“ ( $B_1$ ), „leichte“ ( $B_2$ ), „mittlere“ ( $B_3$ ), „starke“ ( $B_4$ ) Dissonanz), und diese könnten auf zwei verschiedene Arten erzeugt werden (z.B. „einstellungskonträre Information“ ( $A_1$ ), „forced compliance durch Rollenspiel“ ( $A_2$ )). Den entstehenden acht Behandlungskombinationen werden dann zufällig jeweils  $n$  Vpn zugewiesen (siehe Vpl. 6).

		R – UV B (Stärke der Dissonanz)			
		B <sub>1</sub> (keine)	B <sub>2</sub> (leichte)	B <sub>3</sub> (mittlere)	B <sub>4</sub> (starke)
(Vpl. 6)	A <sub>1</sub> (Informa- tion)	Y <sub>111</sub> Y <sub>211</sub> ⋮ Y <sub>n11</sub>	Y <sub>112</sub> Y <sub>212</sub> ⋮ Y <sub>n12</sub>	Y <sub>113</sub> Y <sub>213</sub> ⋮ Y <sub>n13</sub>	Y <sub>114</sub> Y <sub>214</sub> ⋮ Y <sub>n14</sub>
	A <sub>2</sub> (Rollen- spiel)	Y <sub>121</sub> Y <sub>221</sub> ⋮ Y <sub>n21</sub>	Y <sub>122</sub> Y <sub>222</sub> ⋮ Y <sub>n22</sub>	Y <sub>123</sub> Y <sub>223</sub> ⋮ Y <sub>n23</sub>	Y <sub>124</sub> Y <sub>224</sub> ⋮ Y <sub>n24</sub>

Zur Notation:  $Y_{ijk}$  bezeichnet bei solchen Versuchsplänen den Wert der AV Y mit  $i = 1, \dots, s, s', \dots, n$ ;  $j = 1, \dots, l, l', \dots, J$  und  $k = 1, \dots, m, m', \dots, K$ .  $Y_{slm}$  ist also der Wert der beliebig herausgegriffenden Vp s unter der beliebigen Bedingungskombination  $A_l B_m$  oder (AB),..

Noch ein Wort zur Definition der Untersuchungseinheiten („Vpn“). In der Regel sind das tatsächlich einzelne Personen (oder Tiere). Werden jedoch Merkmale ganzer sozialer Gruppen als AV untersucht, stellen diese Gruppen die Untersuchungseinheiten dar. Häufig bezieht sich die AV zwar auf einzelne Personen, diese können aber nicht zufällig den Modalitäten zugewiesen werden, weil sie nur in vorgegebenen Gruppen (z.B. Schulklassen) untersucht werden können. In diesem Fall besteht die einzige Möglichkeit zur Sicherung der internen Validität darin, daß man möglichst viele dieser Gruppen zufällig den Treatmentkombinationen zuordnet - vgl. hierzu insbesondere Glass & Stanley (1970, 501-509) sowie Abschnitt 8.4.5.

Die praktische Durchführung einer Zufallszuordnung erfolgt am besten mit Hilfe von Zufallszahlen. Entsprechende Tabellen finden sich z. B. in Pearson & Hartley (1954, 1972), Fisher & Yates (1963), Edwards (1971), Kreyszig (1973) und Kriz (1978).

Natürlich muß man das Ergebnis einer Zufallszuordnung nicht vorbehaltlos akzeptieren, wenn sich eine offensichtlich „unwahrscheinliche“ Gruppeneinteilung ergeben hat, die die interne Validität stören könnte; dies ist bspw. der Fall, wenn alle männlichen Vpn der Experimental- und alle weiblichen Vpn der Kontrollgruppe zugewiesen worden sind. Nach Möglichkeit ist unter diesen Umständen die Randomisierung zu wiederholen (Mosteller, 1968, 115 f.).

Damit durch Randomisierung die Störfaktoren (IV) ausgeschaltet werden, müssen die dadurch entstandenen Gruppen natürlich bis zur Messung der AV am Ende der Untersuchung erhalten bleiben. Bei längerdauernden Untersuchungen ist das nicht immer der Fall, denn hier kommt es häufiger vor, daß Personen im Laufe der Untersuchung ausscheiden. Hängt diese sog. „experimentelle Mortalität“ mit der UV zusammen, beeinträchtigt sie die interne Validität (Cook & Campbell, 1979). Dies muß immer vermutet werden, wenn eine Behandlungsbedingung für die Probanden unangenehmer ist als eine andere.

Diese Störung der experimentellen Validität ist bei der Interpretation von Untersuchungsergebnissen stets zu berücksichtigen, und zwar unabhängig davon, welches der speziellen statistischen Auswertungsverfahren (siehe Abschn. 10.4.2) gewählt wird.

Bisher haben wir von der Kontrolle individueller Merkmalsunterschiede als Störfaktoren (IV) durch zufällige Zuordnung der Personen (oder anderer Untersuchungseinheiten) zu den Behandlungsgruppen gesprochen. Das Randomisierungsprinzip kann darüber hinaus auch zur Kontrolle anderer Störfaktoren (IV) dienen. Hier nur einige Beispiele für diese sehr wichtige Vorgehensweise:

- Kann die Untersuchung nicht für alle Probanden zum gleichen Zeitpunkt durchgeführt werden, können Unterschiede hinsichtlich der Tageszeit (vor

allem bei allgemeinspsychologischen Untersuchungen) und/oder des Datums der Untersuchung (vor allem in der Sozialpsychologie) Störfaktoren (IV) darstellen. Diese können dadurch ausgeschaltet werden, daß man die zur Verfügung stehenden Untersuchungstermine zufällig den experimentellen Bedingungen zuordnet. Dies erfolgt zum Beispiel, indem man jede ins Laboratorium kommende Versuchsperson zufällig einer Behandlungsgruppe zuweist.

- Wird die Untersuchung mit verschiedenen Versuchsleitern, an unterschiedlichen Orten, mit verschiedenen Apparaten usw. durchgeführt, so sind die Elemente der jeweils zur Verfügung stehenden Menge zufällig den Untersuchungseinheiten oder direkt den Behandlungsgruppen zuzuordnen, damit sie nicht zu Störfaktoren (IV) werden können.

### 3.3.3 Einführung eines Kontrollfaktors

Eine Verallgemeinerung der Konstanthaltung einer potentiellen Störbedingung ist ihre systematische Variation (Bredenkamp, 1969 a). Statt beispielsweise die Darbietungszeit der Reize für alle Probanden konstantzuhalten, kann man verschiedene Darbietungszeiten wählen und diese Variable als sog. „Kontrollfaktor“ (KF) mit  $Q$  Modalitäten in das Design einfügen - siehe Vpl. 7. Abgesehen vom hohen praktischen Aufwand lassen sich natürlich auch gleichzeitig mehrere Kontrollfaktoren einführen, auch zusätzlich zu mehreren bereits vorhandenen experimentellen Faktoren (unabhängige Variablen). Auf jeden Fall werden die Probanden wieder den möglichen Behandlungskombinationen zufällig zugewiesen. In jeder Zelle des Versuchsplanes 7 ergeben sich dann  $n$  Werte auf der AV  $Y$ , nämlich  $Y_{lqm}, \dots, Y_{sqm}, \dots, Y_{nqm}$ , wobei gilt:  $q = 1, \dots, r, \dots, Q$  für den Kontrollfaktor KF und  $k = 1, \dots, m, \dots, K$  für den experimentellen Faktor B.

(Vpl. 7)

		R – UV B (experimenteller Faktor)				
		$B_1$	....	$B_m$	....	$B_K$
R – UV KF (Kontroll- faktor)	$KF_1$					
	$\vdots$					
	$KF_r$			$Y_{srm}$		
	$\vdots$					
	$KF_Q$					

Der Vorteil dieses Vorgehens besteht darin, daß man prüfen kann, inwieweit der Einfluß der UV auf die AV vom jeweiligen Niveau des Kontrollfaktors abhängt (siehe dazu Abschn. 4.3 und 8.5). Zur Kontrolle der Störfaktoren der internen Validität ist die explizite Berücksichtigung des Kontrollfaktors allerdings nicht notwendig. Dazu genügt - um im Beispiel zu bleiben - die Konstanzhaltung der Darbietungszeit oder ihre zufällige Variation über alle Modalitäten der interessierenden UV.

Hier sei schon darauf hingewiesen, daß andere Arten von Kontrollfaktoren eingeführt werden können, um die Populations- oder Situationsvalidität oder die statistische Validität zu erhöhen (s. Abschn. 4.3 und 8.4.1).

### 3.4 Zur Kontrolle der Störfaktoren (IV) bei intraindividueller Bedingungsvariation (Meßwiederholung)

Die Ausführungen im Abschnitt 3.3.1 über die Konstanzhaltung als Methode zur Kontrolle von Störfaktoren (IV) können unmittelbar auf Experimente mit wiederholten Messungen übertragen werden. Wir wollen uns hier nur mit den für wiederholte Messungen spezifischen Störfaktoren (den sog. Sequenzeffekten) beschäftigen. Diese lassen sich nur kontrollieren, indem man die Abfolge der Behandlungen und Messungen variiert. Da bei einer intraindividuellen Variation dieser Reihenfolge (etwa nach der Spiegelbildmethode, siehe Selg, 1975) jede Person unter jeder Behandlungsbedingung mindestens zweimal beobachtet werden muß, können dabei leicht zusätzliche Sequenzeffekte entstehen. Deshalb kommt nur eine interindividuelle Variation der Reihenfolgen in Frage. Dafür gibt es drei Gruppen von Techniken (nach Zimny, 1961, 158-186; Selg, 1975, 50-55; Bredenkamp, 1969a):

#### (1) Zufällige Reihenfolgen

Man bestimmt für jede Untersuchungseinheit die Reihenfolge der Bedingungen zufallsmäßig. Dies geschieht am leichtesten über Tabellen mit Zufallsreihenfolgen einer gegebenen Menge von Zahlen (am ausführlichsten bei Moses & Oakford, 1963; außerdem bei Cochran & Cox, 1957; Fisher & Yates, 1963; Underwood & Shaughnessy, 1975; John & Quenouille, 1977). Dann kann man zwar davon ausgehen, daß man mit zunehmender Zahl der Untersuchungseinheiten der angestrebten Kontrolle der Sequenzeffekte beliebig nahe kommt; besonders bei wenigen Personen können sich aber doch noch beträchtliche Abweichungen ergeben.

## (2) Vollständiges Ausbalancieren

Bei  $K$  Modalitäten der UV sind  $K! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot K$  verschiedene Reihenfolgen möglich. Wird jeder dieser Reihenfolgen eine gleich große Zahl von Versuchspersonen zufällig zugeordnet, so sind alle Sequenzeffekte (auch die höheren) kontrolliert.

## (3) Unvollständiges Ausbalancieren

Bei  $K$  Modalitäten werden  $K$  verschiedene Reihenfolgen so gewählt, daß jede Bedingung gleich häufig an jeder Stelle steht. Schreibt man die  $K$  Reihenfolgen untereinander, nennt man die entstehende Anordnung ein „lateinisches Quadrat“ (Lindquist, 1953). Ist  $K$  eine gerade Zahl, können die Reihenfolgen darüber hinaus so gewählt werden, daß jede Bedingung gleich häufig vor und hinter jeder anderen Bedingung steht. Ist  $K+1$  eine Primzahl, lassen sich Quadrate konstruieren, die auch alle Sequenzeffekte höherer Ordnung kontrollieren (siehe dazu Cochran & Cox, 1957, 133-135; Cox, 1958, 272-274; Alimena, 1962; Edwards, 1971, 227-228; Raghavarao, 1971; Namboodiri, 1972; John & Quenouille, 1977, 196-214; zu den mathematischen Grundlagen siehe Dénes & Keedwell, 1974).

Allgemein können lateinische Quadrate stets verwendet werden, um ausgewogene Kombinationen dreier Faktoren mit der gleichen Anzahl von Modalitäten zu erreichen, wenn diese Ausprägungen aus irgendeinem Grunde nicht vollständig kombiniert („gekreuzt“) werden können (siehe Myers, 1972, 259-281; Winer, 1971, 685-751).

Welche der drei Strategien zur Kontrolle der Sequenzeffekte sollte man anwenden?

Das unvollständige Ausbalancieren stellt die am wenigsten befriedigende Kontrolltechnik dar. Allerdings gibt es zu ihr dann keine Alternative, wenn die Anzahl der Untersuchungseinheiten relativ klein ist oder wenn aus technischen Gründen nur relativ wenige verschiedene Reihenfolgen realisiert werden können.

Das vollständige Ausbalancieren ist nur dann möglich, wenn die Anzahl  $N$  aller Untersuchungseinheiten gleich  $K!$  oder einem ganzzahligen Vielfachen davon ist. Da die Stichprobengröße  $N$  jedoch aufgrund ganz anderer Gesichtspunkte festgelegt werden sollte (siehe Teil 10), dürfte sie nur in Ausnahmefällen gleich  $c \cdot K!$  sein.

Bei der Planung von Untersuchungen mit wiederholten Messungen sollte man daher in der Regel die Reihenfolgen der Behandlungen nach dem Zufallsprinzip bestimmen.



Die genannten Kontrolltechniken sind allerdings nur mit einer sehr wesentlichen Einschränkung wirksam: Durch Variation der Reihenfolgen können Sequenzeffekte nur dann ausgeglichen werden, wenn sie in dem Sinne symmetrisch sind, daß das Ausmaß der von jeder Behandlungsbedingung ausgehenden Sequenzeffekte gleich dem Ausmaß der von anderen auf sie entfallenden Effekte ist (vgl. Zimny, 1961, 165). In vielen Kontexten ist es aber z.B. durchaus wahrscheinlich, daß eine der Behandlungen besonders stark für andere Behandlungen sensibilisiert oder daß unter einer Behandlung der Übungseffekt besonders ausgeprägt ist. Diesen Übungseffekten versucht man oft durch eine ausgedehnte *prä-experimentelle Übungsphase* entgegenzuwirken, durch die erreicht werden soll, daß während des Experiments kaum noch Übungsfortschritte eintreten. Daneben wird die Gefahr von Übertragungseffekten natürlich im allgemeinen um so geringer, je größer die *zeitliche Distanz* der verschiedenen Behandlungen ist (etwa im Drogenbeispiel aus Abschnitt 3.2). Auch ist es zur Vermeidung von Übertragungs- und Sensibilisierungseffekten gerade bei Meßwiederholungen besonders empfehlenswert, den Vpn möglichst wenig Hinweise auf den Zweck der Untersuchung zu geben (vgl. Abschn. 2.5), z.B. indem man in unsystematischer Weise auch andere als die interessierenden unabhängigen Variablen verändert (vgl. Greenwald, 1976, 317).

### 3.5 Versuchspläne mit interindividueller Bedingungsvariation und Vortest

Möchte man auf die Vorteile einer interindividuellen Bedingungsvariation nicht verzichten, besteht aber aus irgendeinem Grund ein Interesse an den Werten der Probanden auf der AV vor der experimentellen Behandlung, kann man die Versuchspläne 4 und 2 wie in Vpl. 8 zu einem Kontrollgruppenexperiment mit Meßwiederholung kombinieren (vgl. Campbell & Stanley, 1963).

(Vpl. 8)	R	$Y_{i0}$	$X_1$	$Y_{i1}$
	R	$Y_{i0}$	$X_2$	$Y_{i2}$

Nachteilig bei diesem Versuchsplan ist, daß die Hypothese ausschließlich an Personen geprüft wird, bei denen in den sog. Vortests die AV bereits einmal vor der experimentellen Behandlung erhoben wurde. Dies setzt die Populationsvalidität des Experiments herab (siehe Abschn. 4.1). Falls man an Vortestwerten interessiert ist, sollte man deshalb den Versuchsplan 8 mit dem Versuchsplan 4 zum sog. „Solomon-Vier-Gruppen-Design“ (s. Vpl. 9) kombinieren (Solomon, 1949; Campbell & Stanley, 1963; Bredenkamp, 1969a; Huck & Sandler, 1973; Oliver & Berger, 1980). Bei diesem Versuchsplan läßt sich auch prüfen, inwieweit die Tatsache, daß ein Vortest durchgeführt worden ist, einen Einfluß auf die Beziehung zwischen der UV und der AV hat (vgl. Abschn. 8.5).

(Vpl. 9)	R	$Y_{i0}$	$X_1$	$Y_{i11}$
	R	$Y_{i0}$	$X_2$	$Y_{i12}$
	R		$X_1$	$Y_{i21}$
	R		$X_2$	$Y_{i22}$

### 3.6 Zur Definition des Experiments und anderer Untersuchungsmethoden

Wir hatten im Abschnitt 3.3 gesehen, daß ohne eine zufällige Zuordnung von Untersuchungseinheiten (Vpn) und experimentellen Bedingungen stets mit einer Störung der internen Validität gerechnet werden muß. Ohne Randomisierung können also Hypothesen über kausale Beziehungen zwischen Variablen nicht streng geprüft werden. Deshalb stellt die Randomisierung die wesentliche Bedingung dafür dar, daß eine Untersuchung als Experiment bezeichnet werden kann (Bredenkamp, 1969a, 1980; Carlsmith, Ellsworth & Aronson, 1976; s.a. unsere unten gegebene Definition).

Wird ein Experiment nicht unter künstlichen („Labor-“) Bedingungen durchgeführt, sondern in „natürlichen“ Situationen, spricht man von einem „*Feldexperiment*“ (s. Bredenkamp, 1969a; Redding, 1970; French, 1972; Kerlinger, 1979; Patry, 1979, 1982; Westmeyer, 1982).

Werden aus irgendeinem Grunde die Vpn *nicht* zufällig den Modalitäten der UV zugeordnet, sprechen Campbell & Stanley (1963) von einem „*Quasi-Experiment*“. Bis zu einem gewissen Grade kann man sich auch bei diesem dem Ideal einer strengen Prüfung annähern, und zwar indem man für möglichst viele relevante Störungshypothesen (IV) nachzuweisen versucht, daß für sie in der konkreten Untersuchungssituation keine entsprechenden Anfangsbedingungen vorliegen (Näheres siehe Cook & Campbell, 1976, 1979).

Alle Untersuchungen, die weder Experimente noch Quasi-Experimente sind, werden als „*Korrelationsstudien*“ bezeichnet. Sie können zwar zur Prüfung von Theorien noch insofern eingesetzt werden, als aus diesen ja stets auch Vorhersagen über korrelative Zusammenhänge ableitbar sind, diese Korrelationen lassen sich in der Regel aber auch durch bewährte Störungshypothesen vorhersagen, so daß keine strenge Prüfung möglich ist. Annäherungen an dieses Ziel sind allenfalls im Rahmen der Prüfung kausalanalytischer Modelle möglich (s. Namboodiri, Carter & Blalock, 1975; Hummell & Ziegler, 1976; Opp & Schmidt, 1976; Kenny, 1979).

Genaugenommen können wir eine Unterscheidung zwischen Experimenten, Quasi-Experimenten und Korrelationsstudien gar nicht für Untersuchungen als Ganzes treffen, sondern nur für einzelne unabhängige Variablen. Häufig finden sich nämlich innerhalb einer Untersuchung neben experimentellen Variablen im strengen Sinne, bei denen die Vpn zufällig den Modalitäten(kombinationen) zugeordnet werden, andere UVn, deren Ausprägungen bei den einzelnen Vpn gar nicht zufällig festgelegt werden können, sondern anderweitig „vorgegeben“ sind (z. B. Geschlecht, Intelligenzquotient, Art der psychischen

Störung).<sup>9)</sup> Kausalaussagen sind in bezug auf derartige Variablen natürlich in aller Regel nicht möglich.

Auf der Grundlage dieser Überlegungen gelangen wir zu folgender Definition der *Experiments*:

*Eine Untersuchung ist bezüglich einer unabhängigen Variablen  $X$  ein Experiment, wenn die gleichen Sachverhalte unter verschiedenen Bedingungen  $X_1, \dots, X_K$  systematisch beobachtet werden und wenn Untersuchungseinheiten und Bedingungen einander zufällig zugeordnet werden bzw. wenn die Reihenfolgen zufällig bestimmt werden, in denen die Untersuchungseinheiten unter diesen Bedingungen beobachtet werden.*

Wir können damit - abweichend von Wundt (1913) - auch Untersuchungen als Experimente auffassen, bei denen die Behandlungsbedingungen nicht völlig *willkürlich* hergestellt werden (können) und/oder bei denen die Bedingungen von anderen Instanzen als dem Experimentator *variiert* werden und/oder die nicht beliebig *wiederholbar* sind. Die entscheidende Frage für die Abgrenzung des Experiments von den anderen angesprochenen Untersuchungsmethoden ist vielmehr, ob eine Randomisierung erfolgt, denn diese stellt eine notwendige Voraussetzung für die strenge Prüfung einer Kausalhypothese dar.

#### 4. Populations- und Situationsvalidität<sup>10)</sup>

##### 4.1 Populationsvalidität (PV)

Unsere Beispielhypothese  $WH_u$  ist in mehrfacher Hinsicht typisch für psychologische Hypothesen. Uns interessieren hier zwei Punkte:

1. Die in der Hypothese vorkommenden theoretischen Begriffe bezeichnen Eigenschaften oder Merkmale, die einzelnen Personen zukommen.
2. Die in der Hypothese getroffene Aussage soll für alle Personen einer unendlichen Population gültig sein.

Beginnen wir bei der Erläuterung dieser beiden Punkte mit dem letzten! Die kognitive Dissonanztheorie enthält keine Einschränkungen ihres Geltungsbe-

---

<sup>9)</sup> Dazu gehören auch Variablen, die sich auf etwas Vergangenes beziehen (z.B. frühkindliche Erfahrungen, Seminarbesuch im vergangenen Semester). Werden solche Variablen mit gegenwärtig beobachtbaren in Beziehung gesetzt, spricht man von „*ex-post-facto*-Untersuchungen“ (s. Meehl, 1970 und Kerlinger, 1979, 579-597).

<sup>10)</sup> Wir vermeiden es, Populations- und Situationsvalidität zusammenfassend als „externe Validität“ zu bezeichnen (Campbell & Stanley, 1963; Bracht & Glass, 1968), weil gerade der Begriff der externen Validität in einer ausgesprochen induktivistischen Weise gebraucht wird (vgl. Gadenne 1976, und Abschn. 1.3).

reichs, d.h. wir müssen davon ausgehen, daß sie den Anspruch erhebt, für alle Personen gültig zu sein. Die aus dieser Theorie von unserem imaginären Forschungsteam abgeleitete Hypothese  $WH_u$  beansprucht zwar nicht unbedingt Allgemeingültigkeit für alle Menschen, wohl aber zumindest für alle Bürger eines bestimmten Staates. Dabei soll sie nicht nur für die endliche Menge der zu einem ganz bestimmten Zeitpunkt lebenden Bürger gelten, sondern auch für die mehr oder minder später existierenden, durch Geburt und Tod, Einwanderung und Fortzug veränderten Populationen. Der Geltungsbereich der Hypothese ist also eine unendliche Menge von Individuen. In der Terminologie der Logik bezeichnete man solche Aussagen als „*unbegrenzte Allsätze*“. Die meisten psychologischen Hypothesen und Gesetze entsprechen derartigen unbegrenzten Allsätzen. Dies gilt auch für solche Hypothesen, die deterministische Aussagen der Form „Wenn . . ., dann . . .“ vermeiden und statt dessen nur Wahrscheinlichkeitsaussagen beinhalten.

Ein Beispiel hierfür ist die folgende Hypothese  $WH_z$ : „Wenn zwischen der Einstellung einer Person zu einem bestimmten Objekt und einem anderen kognitiven Element eine Dissonanz besteht, dann verändert sich mit einer Wahrscheinlichkeit  $q$  die Einstellung dahingehend, daß diese Dissonanz vermindert wird“, oder kurz und formal ausgedrückt: „ $Prob(dE | D) = q$ “.

Die Formulierung von Hypothesen als unbegrenzte Allsätze hat einen entscheidenden Vorteil: Sollte die Hypothese in strengen Prüfversuchen nicht falsifiziert werden und somit als gut bewährte Gesetzesaussage gelten, kann sie zur *wissenschaftlichen Erklärung* einzelner Sachverhalte herangezogen werden. Nach Hempel & Oppenheim (1948; vgl. auch Hempel, 1965; Stegmüller, 1974a; Groeben & Westmeyer, 1975; Küttner, 1979; zur Kritik siehe Suppe, 1977b, 624-632) besteht nämlich die wissenschaftliche Erklärung in der logischen Ableitung der Aussage über den zu erklärenden Sachverhalt (z.B. „Person  $s$  hat ihre Einstellung zum Positiven hin verändert.“) aus einem gut bewährten Gesetz (im Beispiel ist dies unsere  $WH_u$ ) und einigen Anfangsbedingungen (im Beispiel u.a. „Bei Person  $s$  ist Dissonanz erzeugt worden.“). Eine derartige Erklärung kann aber nur dann adäquat sein, wenn das darin verwendete Gesetz eine Aussage ist, die sich auf unendlich viele Anwendungsfälle (d.h. in der Regel: Personen) bezieht (Stegmüller, 1974a). Andernfalls könnte unser exemplarischer Sachverhalt nämlich aus einer Aussage wie „Alle Personen in diesem Raum haben ihre Einstellung zum Positiven hin geändert“ abgeleitet werden, womit sicherlich keine befriedigende wissenschaftliche Erklärung gegeben wäre.

Als *psychologische* Hypothesen oder Gesetze enthalten diese unbegrenzten Allsätze nun Aussagen über einzelne Individuen (im Gegensatz etwa zu soziologischen Hypothesen, in denen zumeist Merkmale von Personengruppen betrachtet werden). Trotzdem können diese Hypothesen nicht für jede einzelne Person falsifiziert werden, da Störungen der internen Validität in aller Regel

nicht mit genügender Sicherheit ausgeschlossen werden können, wenn wir einzelne Personen betrachten, sondern nur wenn mehrere Personen zufällig unterschiedlichen experimentellen Bedingungen zugeordnet werden (s. Abschn. 3.3.2). Die möglichst strenge Prüfung einer Hypothese muß also normalerweise an einer *Gruppe* von Personen erfolgen.

Gehen wir jetzt wieder vom Kriterium einer strengen Prüfung der wissenschaftlichen Hypothese aus! Danach muß eine Untersuchung so beschaffen sein, daß die Wahrscheinlichkeit eines hypothesenkonträren Ergebnisses hoch ist, wenn die Hypothese falsch ist. Im Zusammenhang mit der untersuchten Personengruppe kann das Ziel einer strengen Prüfung in zwei Fällen gefährdet sein: Erstens können keine hypothesenkonträren Ergebnisse eintreten, wenn die in der Untersuchung verwendete Personengruppe gar keine Stichprobe aus der Population darstellt, für die die Hypothese gelten soll (vgl. Bredenkamp, 1979). Wir wollen dies als Störung der *Populationsvalidität erster Art* ( $PV_1$ ) bezeichnen. Sie liegt z.B. dann vor, wenn eine humanpsychologische Hypothese an Tieren oder eine für Schizophrene Gültigkeit beanspruchende Aussage an „Normalen“ überprüft wird.<sup>11)</sup> Zweitens wird eine Untersuchung in aller Regel nur an einer bestimmten Untermenge der Personen (oder anderer Untersuchungseinheiten) durchgeführt, für die die Hypothese gelten soll. Eine Störung der *Populationsvalidität zweiter Art* ( $PV_2$ ) liegt dann vor, falls es eine bewährte Störungshypothese HS, gibt, nach der die Hypothese  $WH_u$  zwar für die Subpopulation gilt, aus der die untersuchten Personen stammen, für andere Untermengen des beanspruchten Geltungsbereiches aber nicht. Dementsprechend ist jede Variable, die Unterpopulationen, für die die zu prüfende Hypothese gilt, von solchen trennt, für die sie nicht gilt, ein möglicher Störfaktor der Populationsvalidität zweiter Art (vgl. Gadenne, 1976).

Beispiele: Die Gültigkeit einer wissenschaftlichen Hypothese kann beschränkt sein auf Subpopulationen von Freiwilligen (Rosenthal & Rosnow, 1969b), von Probanden, die mit psychologischen Experimenten vertraut sind, von Probanden, die motiviert genug sind, ein langdauerndes Experiment durchzuhalten, von Personen mit hoher Bewertungsangst (Weber & Cook, 1972), von Arbeitern eines bestimmten Betriebes, von Personen, die durch Messung der AV vor der Behandlung sensibilisiert worden sind (Lana, 1969) usw. In all diesen Fällen ist die PV gestört, wenn eine Untersuchung nur mit Personen aus der jeweiligen Unterpopulation durchgeführt wird.

*Zusammenfassend können wir also formulieren: Eine Untersuchung zur Prüfung einer Kausalhypothese ist um so strenger, je weniger sie sich auf Probanden aus bestimmten Untermengen der Population beschränkt, für die die Hypothese gelten soll.*

---

<sup>11)</sup> Oft sind Untersuchungen an Personen aus dem Gültigkeitsbereich der Hypothese aus technischen oder ethischen Gründen gar nicht durchführbar - man denke nur an die Erprobung neuer Medikamente. Werden in einem solchen Fall Tiere als Modelle menschlicher Organismen benutzt, so kann in diesen Untersuchungen die höchste zu diesem Zeitpunkt realisierbare Populationsvalidität gegeben sein.

Die untersuchten Probanden brauchen aber keine Zufallsstichprobe aus irgendeiner (Sub-)Population zu sein, denn jede Kausalhypothese kann (weil sie sich - wie begründet - notwendigerweise auf eine *unendliche* Population von Anwendungsfällen bezieht) an *beliebigen* Teilgruppen dieser Population falsifiziert werden (Holzkamp, 1964; Bredenkamp, 1972; Gadenne, 1976; s.a. Abschn. 8.2.6).

## 4.2 Situationsvalidität (SV)

Bei der Formulierung wissenschaftlicher Hypothesen und Theorien sollte stets angegeben werden, für welche raum-zeitlichen Bedingungskonstellationen (kurz: für welche Situationen) sie gültig sein sollen (Westmeyer, 1982). Dieser für eine Theorie beanspruchte Geltungsbereich wird meist aus einer sehr großen, in der Regel sogar unendlichen Menge von Situationen bestehen. Im Zusammenhang mit der bei einer empirischen Prüfung einer Theorie vorliegenden Situation können wir zwei Fälle unterscheiden, in denen das Ziel einer strengen Prüfung gefährdet ist: erstens wenn die Hypothese oder Theorie gar nicht für die Situation gelten soll, in der der Prüfversuch durchgeführt wird (Störung der SV erster Art); zweitens wenn die Vermutung berechtigt ist, daß die zu prüfende Hypothese zwar unter den gegebenen Umständen erfüllt ist, nicht aber unter (einigen oder vielen) anderen Kombinationen von situationalen Bedingungen, für die ihre Gültigkeit ebenfalls beansprucht wird (Störung der SV zweiter Art). Entsprechend der Definition von Störfaktoren der Populationsvalidität wollen wir eine Variable als Störfaktor der Situationsvalidität - Störfaktor (SV) - bezeichnen, wenn sie nach einer bewährten Störungshypothese Situationen definiert, in denen die zu prüfende Hypothese nicht gilt.

Störfaktoren der Situationsvalidität können von sehr unterschiedlicher Art sein. So mag eine Hypothese  $WH_v$  nur gültig sein, wenn den Probanden bewußt ist, daß sie an einer Untersuchung teilnehmen (Hawthorne- oder Placebo-Effekt, vgl. auch den von Klauer (1973, 558) beschriebenen „Novitätseffekt“ in der pädagogischen Forschung), wenn die abhängige Variable auch bereits vor der Behandlung gemessen wurde (Vortest-Sensitivierung), wenn die AV sofort nach der Behandlung erfaßt wird (mangelndes Überdauern des Effekts), wenn die Prüfung durch Versuchsleiter mit bestimmten Eigenschaften und Verhaltensweisen, in Räumen eines Forschungsinstituts, unter den Bedingungen eines Laborexperiments, unter Verwendung ganz bestimmter Apparaturen, bei einer bestimmten Dauer der experimentellen Behandlung, zu einer bestimmten Tageszeit, zeitlich kurz nach einer bestimmten Fernsehsendung usw. stattfindet (vgl. Bracht & Glass, 1968). Solche und viele ähnliche Faktoren können die Validität einer Untersuchung einschränken.

*Daraus ergibt sich allgemein: Eine Untersuchung zur Prüfung einer Kausalhypothese ist um so strenger, je weniger sie sich auf ganz bestimmte zeitliche, räumliche und situationale Umstände aus dem Geltungsbereich der Hypothese beschränkt.*

Umfaßt der Geltungsbereich einer Hypothese nicht nur Laborsituationen, müssen im Interesse einer möglichst strengen Prüfung dieser Hypothese also auch Feldexperimente durchgeführt werden (vgl. Abschn. 3.6 und 5).

### 4.3 Zur Kontrolle der Störfaktoren (PV und SV)

Liegt eine Vermutung darüber vor, daß eine Variable  $V$  ein Störfaktor der Populations- oder Situationsvalidität ist, kann man dies stets prüfen, indem man die Untersuchung unter mindestens zwei Ausprägungen dieser Variablen durchführt, diese Variable also als Kontrollfaktor einführt (vgl. Abschn. 3.3.3). Betrachten wir dazu ein Beispiel! In einem Experiment zur Prüfung unserer Hypothese  $WH_u$  wird die experimentelle Bedingung „keine Dissonanz“ ( $A_1$ ) der Bedingung „Dissonanz“ ( $A_2$ ) gegenübergestellt, und es interessiert uns, inwieweit die Zugehörigkeit zu einer bestimmten Berufsgruppe (Faktor B) die Resultate beeinflußt. Faktor B umfasse drei Modalitäten: „Student“ ( $B_1$ ), „Beamter“ ( $B_2$ ) und „freier Beruf“ ( $B_3$ ). Es seien nun unter den 6 Bedingungskombinationen folgende Mittelwerte der AV aufgetreten (je höher die Werte sind, desto positiver ist die Einstellung):

Tabelle 4.1:

	$B_1$	$B_2$	$B_3$
$A_1$	0	5	10
$A_2$	2	4	6

Offensichtlich ist die Vorhersage der Hypothese  $WH_u$  bei Studenten eingetreten, jedoch nicht bei Beamten und Freiberuflern. Nach unserer Definition ist B also ein Störfaktor der Populationsvalidität. In statistischer Terminologie spricht man davon, daß eine „*disordinale Interaktion*“ bezüglich des Faktors A vorliegt (siehe im einzelnen Abschn. 8.5).

Dies gilt allgemein: *Eine Variable ist genau dann ein Störfaktor der Populations- oder Situationsvalidität, wenn bezüglich des Behandlungsfaktors eine disordinale Interaktion besteht.* Wird eine derartige Wechselwirkung wiederholt festgestellt, sollte zumindest der Geltungsbereich der Hypothese eingeschränkt werden. Wie man feststellen kann, ob zwei Variablen interagieren und ob ggf. diese Interaktion disordinal ist, werden wir im Abschnitt 8.5 noch ausführlicher besprechen.

Innerhalb einer Untersuchung wird sich stets nur eine sehr kleine Zahl von potentiellen Störfaktoren der SV oder PV auf diese Weise untersuchen lassen. Zur Klärung der Frage, in welchem Ausmaß die Strenge der Prüfung einer

Hypothese durch solche Faktoren eingeschränkt ist, muß man sich deshalb zum einen auf entsprechende Ergebnisse aus Untersuchungen zur Prüfung anderer Hypothesen stützen, zum anderen kann eine einigermaßen fundierte Beurteilung der Gültigkeit einer Kausalhypothese nur erfolgen, nachdem sie durch mehrere Untersuchungen in verschiedenen Subpopulationen und Situationen überprüft wurde. Da aus den unendlichen Mengen der Personen und Situationen, für die die Hypothese gelten soll, keine repräsentativen Stichproben zu ziehen sind, wird das Ziel einer strengen Prüfung der Hypothese durch ein Forschungsprogramm aus mehreren Untersuchungen am besten erreicht, wenn aus dem Geltungsbereich Personengruppen und Situationen bewußt so ausgewählt werden, daß sie möglichst verschiedenartig sind (vgl. Cook & Campbell, 1979, 74-80; s.a. Dipboye & Flanagan, 1979) und daß in anderen Kontexten bewährte oder (aus theoretischen Überlegungen) besonders plausible Störungshypothesen direkt geprüft werden können.

### *5. Beziehungen zwischen den Validitätsarten*

Zwischen interner Validität auf der einen und Populations- und Situationsvalidität auf der anderen Seite besteht im allgemeinen eine gegenläufige Beziehung. Zunächst einmal sind Störungen der SV und PV häufig bedingt durch Maßnahmen zur Kontrolle möglicher Störfaktoren der internen Validität. Ein Beispiel möge dies verdeutlichen: Die IV ist gefährdet, wenn jede Experimentalgruppe einen anderen Versuchsleiter hat. Diese Gefahr kann z.B. dadurch ausgeräumt werden, daß man für alle Gruppen das Experiment durch den gleichen VL durchführen läßt, die Variable „Versuchsleiter“ also konstant hält (vgl. Abschn. 3.3.1). Dann besteht aber die Gefahr, daß das erhaltene Ergebnis spezifisch ist für Experimente unter Leitung dieser Person und daß sich bei anderen VL mit anderen Eigenschaften andere Resultate eingestellt hätten. Durch Konstanthaltung von Bedingungen kann also die Situationsvalidität eingeschränkt werden. Umgekehrt wird durch eine Erhöhung der SV oft die interne Validität herabgesetzt. So ist unter natürlich auftretenden situationalen Bedingungen die Konstanthaltung potentieller Störfaktoren (IV) oder die Randomisierung der Untersuchungseinheiten, -zeitpunkte und -räume meist wesentlich schwieriger. Oft sind diese Maßnahmen gar nicht durchzuführen, so daß allenfalls Quasi-Experimente möglich sind. Wegen dieser Gegenläufigkeit der betrachteten Validitätsaspekte sollte eine Entscheidung über eine Hypothese, deren Geltungsbereich auch „natürliche“ Situationen umfaßt, erst aufgrund der Ergebnisse von mehreren Untersuchungen Verschiedenster Art getroffen werden.

Läßt sich eine Hypothese jedoch tatsächlich nur in Laborsituationen mit einer ausreichenden internen Validität überprüfen, muß ihr Geltungsbereich strenggenommen auf diese Situationen beschränkt werden. Aussagen über andere



Situationen lassen sich dann zwar logisch nicht ableiten, es spricht jedoch nichts dagegen, aus Theorien und Hypothesen, die sich in strengen Prüfungen *gut bewährt* haben, sog. *technologische Prognosen* zu gewinnen (Gadenne, 1976; Brocke, 1978, 1979). Mit Hilfe von Gesetzmäßigkeiten, die (u.U. nur im Labor) relativ gut gesichert worden sind, werden also Voraussagen über Variablenzusammenhänge in komplexen praktischen Situationen wie z.B. einer Psychotherapie gemacht. Anders als bei der isolierten Bedingungsvariation des intern validen Experiments sind dort aber ganz sicher auch andere Variablen wirksam.

Im Zusammenhang mit der Variablenvalidität hatten wir im Abschnitt 2.5 erwähnt, daß mit der UV des Experiments Aspekte des Versuchsleiterverhaltens und Hypothesen der Versuchspersonen über die von ihnen erwarteten Verhaltensweisen konfundiert sein können und daß dadurch fälschliche Falsifikationen und Bestätigungen entstehen können. Wir erinnern hier an diesen Punkt, weil die Höhe der Gefahr einer solchen Konfundierung abhängig sein kann von der Situation, in der das Experiment durchgeführt wird. Je „unauffälliger“ und „natürlicher“ die Herstellung der Behandlungsbedingungen und die Registrierung der abhängigen Variablen erfolgt, je weniger Hinweise auf Hypothesen und Erwartungen des Experimentators die Situation den Probanden gibt, desto geringer ist die Gefahr derartiger Störungen der Variablenvalidität.

Von den möglichen Maßnahmen zur Vermeidung solcher Forschungsartefakte seien nur einige der wichtigeren genannt:

- Die Täuschung der Probanden über ihre Rolle als Versuchspersonen und/oder über die Erwartungen des Versuchsleiters;
- Durchführung von *Feldexperimenten* und Simulationsstudien (Rollen-spiele);
- Erfassung der abhängigen Variablen durch *nicht-reaktive* Methoden (vgl. auch Abschnitt 2.5; zu weiteren Einzelheiten siehe etwa Bredenkamp, 1969a; Summers, 1970; Bungard & Lück, 1974; Webb et al., 1975; Carls-mith, Ellsworth & Aronson, 1976).

Wir wollen nun die Beziehung zwischen der Variablenvalidität und den anderen Arten der experimentellen Validität genauer betrachten. Diese Beziehung hängt ganz entscheidend von der zu prüfenden Hypothese ab (Gadenne, 1976, 99-103). Bezieht sich die Hypothese auf ideale, künstliche Situationen (z.B. weil die Annahme der Konstanz aller anderen Bedingungen notwendig ist), ist es meist relativ einfach, gleichzeitig alle Teilaspekte der experimentellen Validität in befriedigendem Ausmaß sicherzustellen. Anders ist es bei Theorien und Hypothesen, deren Begriffe nur in natürlichen Situationen realisiert werden können, wie z. B. bei vielen entwicklungspsychologischen Theorien oder bei Hypothesen über das Verhalten in sozialen Gruppen. Sollen Prüfungen

solcher Hypothesen eine hohe Variablenvalidität haben, müssen sie in Situationen durchgeführt werden, die häufig nur eine ungenügende Kontrolle der Störfaktoren der internen Validität erlauben. Hier stellen also Variablenvalidität und interne Validität konträre, wenn nicht sogar unvereinbare Ziele dar. Typisch für die Psychologie sind jedoch Hypothesen und Theorien, die sowohl im Labor wie in natürlich auftretenden Situationen Gültigkeit beanspruchen. Diese können zwar durchaus im Labor-Experiment einer Prüfung unterzogen werden; wie schon im Zusammenhang mit der Situationsvalidität erwähnt, erfordert eine strenge Prüfung dieser Hypothesen aber auch ihre Überprüfung in anderen Umgebungen, also z.B. die Durchführung von Feld- oder Quasi-Experimenten.

Insgesamt gesehen sind Variablenvalidität und interne Validität von grundlegender Bedeutung als Populations- und Situationsvalidität. Bei unzureichender Variablenvalidität oder interner Validität ist ein Experiment auf keinen Fall eine strenge Prüfung der betrachteten Kausalhypothese, bei mangelhafter Populations- oder Situationsvalidität dagegen ist immerhin noch eine strenge Prüfung der in ihrer Gültigkeit auf die entsprechende Population oder Situation eingeschränkten Hypothese möglich.

## 6. Statistische Validität

Wir haben im Abschnitt 4.1 gesehen, daß psychologische Hypothesen typischerweise als Kausalaussagen formuliert sind, die für jede einzelne Person (oder Gruppe oder Schulklassse etc.) einer bestimmten Population gelten sollen. Eine empirische Überprüfung derartiger Hypothesen wäre dann am strengsten, wenn man ihre Gültigkeit für jede einzelne Person (oder Gruppe etc.) untersuchen könnte. Nun gibt es zwar durchaus Methoden zur Einzelfallanalyse (siehe etwa Hersen & Barlow, 1976; Kratochwill, 1978; Petermann & Hehl, 1979; Petermann, 1981), doch ist für eine hinreichend strenge Prüfung von Kausalhypothesen in aller Regel die Zusammenfassung mehrerer untersuchter Personen (oder allgemein: mehrerer experimenteller Untersuchungseinheiten) notwendig, und zwar in erster Linie aus den folgenden zwei Gründen, die sich aus unseren bisherigen Ausführungen ergeben (vgl. auch Gadenne, 1976, 88f.):

- (1) Die interne Validität einer Untersuchung kann nur in ausreichendem Maße gesichert werden, wenn mehrere experimentelle Einheiten den Ausprägungen der unabhängigen Variablen zufällig zugeordnet werden (vgl. Abschnitt 3.3.2; bei Meßwiederholungen entsprechend - siehe Abschnitt 3.4).
- (2) Ordnen wir einem theoretischen Begriff eine empirische Variable zu, so ist diese i. a. keine eindeutige oder fehlerfreie Entsprechung des theoretischen Begriffs.

schen Begriffs. Dies wird am Beispiel leicht deutlich: Das Ergebnis eines Probanden in einem Intelligenztest kann man nicht als fehlerfreies Maß für die Ausprägung des theoretischen Begriffs „Intelligenz“ ansehen, weil das verfügbare „Hintergrundwissen“ die Information enthält, daß es z.B. auch durch den augenblicklichen Ermüdungszustand beeinflusst wird. Hat daher eine Person ein unterdurchschnittliches Ergebnis in diesem Test erzielt, so werden wir sie trotzdem nicht sofort als „unterdurchschnittlich intelligent“ bezeichnen, wenn wir z.B. wissen, daß dieser Test nach einer anstrengenden Nachtschicht durchgeführt worden ist. Zur Notwendigkeit derartiger „Ausweichklauseln“ und ihren Konsequenzen für die Zuordnung von theoretischen und empirischen Begriffen siehe insbesondere Herrmann (1973).

Beobachtet man - um beim einfachsten Fall zu bleiben - jeweils  $n$  Versuchspersonen unter zwei experimentellen Bedingungen, wird man praktisch hinsichtlich aller in Psychologie und anderen Sozial- und Biowissenschaften interessierenden abhängigen Variablen feststellen, daß auch bei Personen aus der gleichen Behandlungsbedingung verschiedene Werte auftreten. Diese Variation entsteht sowohl durch Unterschiede zwischen den Probanden hinsichtlich der verschiedensten Persönlichkeitsmerkmale (z. B. Alter, Geschlecht, Intelligenz) und hinsichtlich der vorausgegangenen Erfahrungen (z.B. mit ähnlichen Aufgaben) als auch durch - nicht ganz vermeidbare - Unterschiede in den Durchführungsbedingungen bei den verschiedenen Probanden sowie durch geringe Zuverlässigkeit der Meßinstrumente und/oder Beobachter.

Betrachten wir wieder unsere Beispielhypothese  $WH_u$ , aus der ja die empirisch prüfbare Aussage folgt, daß unter der (Dissonanz-)Bedingung  $X_2$  die Einstellungsvariable  $Y$  größer ist als unter der (Kontroll-)Bedingung  $X_1$ . Auch angesichts der stets anzutreffenden Fehlervarianz der Variablen  $Y$  wäre die Prüfung dieser Hypothese völlig unproblematisch, wenn bei intraindividuellem Bedingungsvariation alle Individuen unter  $X_2$  einen höheren  $Y$ -Wert hätten als unter  $X_1$  bzw. wenn bei interindividuellem Bedingungsvariation die  $Y$ -Werte aller unter  $X_2$  beobachteten Personen größer wären als bei den Personen unter  $X_1$ . Leider ist das in der Psychologie (und in verwandten Wissenschaften) praktisch nie der Fall, vielmehr überlappen sich die Verteilungen der Werte auf der abhängigen Variablen unter den verschiedenen Bedingungen mehr oder minder stark. Deshalb lassen sich hier Kausalhypothesen nur dadurch überprüfen, daß man aus ihnen *statistische Hypothesen* ableitet.<sup>12)</sup> Wir wollen dies etwas näher erläutern (nach Bredenkamp, 1972): Aus  $WH_u$  folgt beispielsweise die statistische Hypothese, daß der Mittelwert der Verteilung der  $Y$ -Werte unter

<sup>12)</sup> Der Gedanke einer implikativen Beziehung zwischen wissenschaftlichen statistischen Hypothesen geht u.W. auf Meehl (1967) zurück und wurde von Bredenkamp (1969b, 1972, 1979, 1980) aufgegriffen und weiterentwickelt.

der Bedingung  $X_2$  größer ist als der entsprechende Mittelwert unter  $X_1$ , oder kurz  $H_1: \mu_2 > \mu_1$  bzw. anders geschrieben  $\mu_2 - \mu_1 > 0$ .<sup>13)</sup> Logisch gesehen besteht zwischen der wissenschaftlichen Hypothese  $WH_u$  und der statistischen Hypothese  $H_1$  eine Implikationsbeziehung:  $WH_u \rightarrow H_1$ .<sup>14)</sup> Diese Beziehung eröffnet die Möglichkeit, wissenschaftliche Hypothesen über die Prüfung von statistischen Hypothesen zu falsifizieren. Denn kann man feststellen, daß die statistische Hypothese falsch ist, muß bei Gültigkeit der Implikationsbeziehung auch die übergeordnete wissenschaftliche Hypothese falsch sein. Zeigt sich dagegen, daß die statistische Hypothese richtig ist, hat sich auch die wissenschaftliche Hypothese in dem betr. Experiment bewähren können.

Diese Verfahrensweise scheitert in der dargestellten einfachen Form jedoch daran, daß statistische Hypothesen selbst grundsätzlich weder beweisbar noch widerlegbar sind, weil sie Aussagen über unendliche Populationen enthalten (Bredenkamp, 1972). Man kann lediglich Kriterien festlegen, die Angaben über die Bedingungen enthalten, unter denen man sich für die Annahme oder für die Ablehnung einer statistischen Hypothese entscheiden will (Gadenne, 1976, 86).

Auch wenn man aufgrund der erhobenen Daten zu der Entscheidung gelangt, daß die  $H_1$  falsch ist, kann man wegen der Unsicherheit über den Wahrheitswert von statistischen Hypothesen daraus nie logisch die Falschheit der übergeordneten Kausalhypothese  $WH_u$  ableiten. Man kann lediglich eine weitere methodologische Regel akzeptieren und befolgen, nach der die Entscheidung für die Falschheit von  $H_1$  notwendige Voraussetzung für eine Falsifikation der Kausalhypothese  $WH_u$  ist und nach der die Kausalhypothese solange als bewährt gelten kann, wie man die von ihr implizierte statistische Hypothese für richtig hält.

Unabhängig von der Wahl des Verfahrens, mittels dessen man zu Entscheidungen über das Zutreffen der statistischen Hypothesen gelangen will, besteht stets die Möglichkeit, daß falsche Entscheidungen getroffen werden - wir kommen auf dieses Problem im Abschnitt 7.3 zurück.

<sup>13)</sup> Zur genaueren Definition des Begriffs der statistischen Hypothese siehe Abschn. 7.1. Betont sei auch, daß aus der  $WH_u$  noch andere statistische Hypothesen ableitbar sind, etwa über Mediane („ $Md_2 > Md_1$ “) oder über die stochastische Größe von Zufallsvariablen („ $F_1(Y) > F_2(Y)$  für alle Werte der AV  $Y$ “).

<sup>14)</sup> Eine Implikationsbeziehung zwischen wissenschaftlicher und statistischer Hypothese kann auch dann hergestellt werden, wenn die wissenschaftliche Hypothese eine Wahrscheinlichkeitsaussage der Form  $\text{Prob}(dE \mid D) = q$  ist (vgl. Abschnitt 4.1). Auch wenn man Theorien als axiomatisch formulierte abstrakte Strukturen betrachtet (s. Groeben & Westmeyer, 1975, 71-75; Westmeyer, 1981), kann die Angemessenheit einer solchen Theorie für einen bestimmten empirischen Sachverhalt geprüft werden, indem man aus den Axiomen statistische Hypothesen ableitet und diese überprüft (vgl. Westermann, 1980).

Das Kriterium der strengen Prüfung von Kausalhypothesen erfordert nun, vereinfacht ausgedrückt (vgl. Abschnitt 1.3), daß die Wahrscheinlichkeit hypothesenkonträrer Ergebnisse hoch ist, wenn die Kausalhypothese falsch ist, und daß sie niedrig ist, wenn die Hypothese „wahr“ ist.

*Daraus folgt: Die Prüfung einer Kausalhypothese über eine statistische Hypothese ist (alle anderen Bedingungen als konstant vorausgesetzt) um so strenger, je geringer die Wahrscheinlichkeiten für Fehler bei der Entscheidung über die abgeleitete statistische Hypothese sind.*

Diejenigen Faktoren, die diese Fehlerwahrscheinlichkeiten erhöhen, wollen wir als „Störfaktoren der statistischen Validität (StatV)“ bezeichnen (in Anlehnung an Cook & Campbell, 1976). Wodurch die statistische Validität einer Untersuchung konkret beeinträchtigt werden kann und wie man diese Störungen zu vermeiden versuchen kann, werden wir im Teil 8 erläutern. Grundlage dafür ist der Teil 7, in dem die für unser Thema wichtigsten Aspekte des statistischen Hypothesentestens dargestellt sind. Die Teile 9 und 10 sind dann zwei Teilaspekten der statistischen Validität gewidmet: den Maßen für die Größe „experimenteller Effekte“ und der nach bestimmten Kriterien optimalen Wahl der Zahl N von Untersuchungseinheiten. Die Überlegungen in den Teilen 7 bis 10 führen uns zu Empfehlungen, wie man bei der Planung und Durchführung von statistischen Hypothesenprüfungen vorgehen sollte und wie man auf dieser Basis zu Entscheidungen über Falsifikation oder Beibehaltung wissenschaftlicher Hypothesen gelangen kann. Diese Planungs- und Entscheidungsstrategie ist im abschließenden Teil 11 dargestellt.

## 7. Eine Strategie zur Entscheidung zwischen statistischen Hypothesen: Der Signifikanztest

### 7.1 Überblick über verschiedene alternative Strategien

Die Entscheidung über das Zutreffen von statistischen Hypothesen, die i.a. über Verteilungen von Zufallsvariablen oder deren Parameter formuliert werden, kann grundsätzlich auf mehrere Arten erfolgen, die wir im folgenden nach Menges (1972) und Barnett (1973) kurz und summarisch aufführen.<sup>15)</sup>

1. Das Verfahren nach Bayes setzt weitgehende Kenntnisse über die o.a. Verteilungen voraus, die häufig nur unter Verwendung subjektiver Annahmen verfügbar gemacht werden können - siehe zu diesem Modell im einzelnen Edwards, Lindman & Savage (1963), Menges (1972, 272-274), Philips (1974), Ruppell (1977), Rützel (1979, 1980) sowie Kleiter (1981).

<sup>15)</sup> Wir greifen die hier nicht erläuterten Begriffe im folgenden auf.

2. Das von Fisher (e. g. 1935) entwickelte Fiduzial-Modell ist anwendbar ohne Vorkenntnisse über die Verteilungen von Zufallsvariablen, dabei „aber an strenge Bedingungen geknüpft“ (Menges, 1972, 275). Es ist wenig gebräuchlich, hat sich aber als das bislang einzig exakte Modell zur Lösung des sog. „Behrens-Fisher-Problems“ erwiesen (vgl. dazu Abschnitt 8.2.4.1). Eine kurze Darstellung des Fiduzial-Modells findet man etwa bei Menges (1972, 275-279).
3. Das ebenfalls auf Fisher zurückgehende Likelihood-Modell erlaubt die Bestimmung der (relativen) „Plausibilität“ (Likelihood) von Parametern - siehe dazu im einzelnen u.a. Edwards (1972), Menges (1972, 279-283) und Witte (1980, 34-40).
4. Das Konfidenz-Modell ermöglicht die Angabe von Zufallsintervallen, die mit vorgegebener Wahrscheinlichkeit einen bestimmten Parameter umschließen oder überdecken. Die weite Verbreitung dieses Modells dokumentiert sich auch darin, daß es in fast jedem Lehrbuch der Statistik vglw. ausführlich erörtert wird; siehe zur Einführung etwa Menges (1972, 282-286), Hays (1977, Kap. 9) oder Witte (1980, 23-26).
5. Vor dem Hintergrund dieses Konfidenz-Konzepts wurden die „klassischen“ Theorien des Signifikanztests entwickelt, nämlich das Modell von Fisher (e.g. 1925, 1956) und das Modell von Neyman & Pearson (1933a, b, 1936, 1938; Neyman, 1952).
6. Das Modell der sequentiellen Tests von Wald (1947) geht über diese Ansätze hinaus und kann als Spezialfall einer allgemeinen Theorie der statistischen Entscheidungsfunktionen aufgefaßt werden (Wald, 1950); zur Darstellung siehe etwa Weber (1967, 395-482) und Wetherill (1975).

Vergleiche bzgl. der Leistungsfähigkeit einiger der aufgeführten Ansätze findet man etwa bei Bredenkamp (1972, 134-150), Barnett (1973) und Witte (1977, 1980); zur Verbindung verschiedener Verfahren zu einem mehrstufigen Inferenzmodell siehe Witte (1980).

Üblicherweise wird von Psychologen, Pädagogen und Sozialwissenschaftlern der Signifikanztest zur Entscheidung über das Zutreffen von statistischen Hypothesen herangezogen - vgl. zu dieser Behauptung u.a. Sterling (1959), Cohen (1962), Edgington (1964 a, 1974), Smart (1964), Bozarth & Roberts (1972), Bredenkamp (1972, 9) und Witte (1980, 17). Aus diesem Grunde beschränken wir unsere Ausführungen auf dieses Verfahren.

Dabei ist anzumerken, daß es „den“ Signifikanztest nicht gibt. Vielmehr bestehen in Theorie und Praxis divergierende Auffassungen über das Rationale, das der Gruppe von Verfahren zugrunde liegt, die unter der Bezeichnung „Signifikanztest“ zusammengefaßt werden, und über ihre adäquate Anwendung.

Die theoretische Kontroverse entstand in den dreißiger Jahren, als Jerzy Neyman und Egon S. Pearson (1933 a, b, 1936, 1938) das hauptsächlich in den

zwanziger Jahren von Sir Ronald A. Fisher (1925, 1956) entwickelte Modell der statistischen Hypothesenprüfung durch ihre Theorie der Fehler 1. und 2. Art zu erweitern suchten und Fisher diese Erweiterung strikt zurückwies. Wir können aus Raumgründen auf diese Kontroverse nicht näher eingehen, sondern müssen den Interessenten auf zusammenfassende Darstellungen verweisen, etwa auf den Reader von Morrison & Henkel (1970) sowie die Veröffentlichungen von Spielman (1974, 1978), Carlson (1976), Chase & Tucker (1976) sowie Witte (1980).

Ebensowenig können wir die Grundgedanken dieser Testtheorien hier vorstellen; die in den nachfolgenden Abschnitten enthaltenen skizzenhaften Darstellungen einiger für das Verständnis späterer Ausführungen notwendiger Aspekte des Signifikanztests setzen daher die Kenntnis der Grundgedanken voraus. Sofern der Leser mit diesen nicht vertraut ist, mag er sich in einem der zahlreichen Lehrbücher informieren, etwa in den eher mathematisch oder wissenschaftstheoretisch orientierten Einführungen von Kendall & Stuart (1961), Hacking (1965, 1976), Fisz (1970), Menges (1972), Stegmüller (1973a, b) oder Mood, Graybill & Boes (1974); eher für Psychologen und Sozialwissenschaftler verfaßte Texte stammen von Hays (1963, 1977), Stilson (1966), Kriz (1978), Leiser (1978), Bortz (1979), Haagen & Seifert (1979) sowie Witte (1980) - diese Zusammenstellung erhebt selbstverständlich keinen Anspruch auf Vollständigkeit.

Gegen bestimmte Aspekte der Anwendung von Signifikanztests im Bereich der Psychologie und verwandten Wissenschaften sind wiederholt und zu Recht teils schwerwiegende Bedenken ins Feld geführt worden (vgl. zu einigen Einzelheiten etwa die Abschnitte 7.4.1 und 7.4.2). Diese haben Autoren wie etwa Kleiter (1969), Harnatt (1975, 1979), Derrick (1976), Guttman (1977) und Carver (1978) zu der Empfehlung veranlaßt, auf Signifikanztests weitestgehend oder völlig zu verzichten.

Wir sind zwar mit Tukey (1977) der Auffassung, daß etwa einer exploratorischen Datenanalyse (EDA) im Forschungsprozeß eine wesentlich gewichtigere Rolle beigemessen werden muß, als es derzeit geschieht, meinen aber andererseits mit Bredenkamp (1972, u.a. 134-150; 1980) und Witte (1980), daß ein „richtig angewendeter“ Signifikanztest trotz aller immanenten Schwächen (derzeit noch) unverzichtbar ist.

Wenden wir uns nun einigen Aspekten der Prüfung von statistischen Hypothesen mittels Signifikanztests zu!

## 7.2 Kurzer Abriß einiger Charakteristika von Signifikanztests

Mit dem Terminus „statistische Hypothese“ belegt man jede Annahme über die Wahrscheinlichkeits- oder die (theoretische) Populationsverteilung einer beobachtbaren Zufallsvariablen oder aber über einen oder mehrere Parameter (vgl. Kendall & Stuart, 1961, 161; Hays, 1977, 335; Haagen & Seifert, 1979, 167-175).

Der Terminus „Parameter“ bezeichnet dabei eine endliche Menge von Konstanten, durch die eine Populationsverteilung näher spezifiziert wird - siehe zu Einzelheiten etwa Menges (1972, 214-216); zum hier nicht erläuterten Terminus „Zufallsvariable“ siehe u.a. Stilson (1966, 121-125), Menges (1972, 139-145) oder Hays (1977, 110-133).

Als Beispiel für eine Populationsverteilung sei die Gruppe der sog. „Normalverteilungen“ genannt, die durch eine einzige Verteilungsfunktion charakterisiert ist (siehe etwa Hays, 1977, 297); die einzelnen Vertreter dieser Gruppe unterscheiden sich darin, welche numerischen Werte die beiden Parameter dieser Verteilung annehmen, nämlich die „Varianz  $\sigma^2$ “ und der „Mittel-“ oder „Erwartungswert  $\mu$ “ der entsprechenden Zufallsvariablen  $Y$ .

Könnte man die theoretisch möglichen Realisierungen der Zufallsvariablen  $Y$ , also die Gesamtmenge der Daten, im konkreten Fall vollständig untersuchen, wäre eine sichere Entscheidung über das Zutreffen oder Nicht-Zutreffen der statistischen Hypothese möglich.

Der empirisch arbeitende Wissenschaftler ist jedoch fast stets darauf angewiesen, weniger sichere und weniger definitive Aussagen aufgrund einer nur beschränkten Menge von Daten zu treffen. Diese Untermenge der Population von Daten heißt „Stichprobe (von Daten)“; sie umfaßt eine mögliche Auswahl aus der Grundgesamtheit.

Man spricht von einer „Zufallsstichprobe“, wenn alle möglichen Auswahlvarianten der gleichen Größe  $n$  die gleiche Wahrscheinlichkeit haben, als Stichprobe zu fungieren (vgl. Fisz, 1970, 394; Hays, 1977, 72); über andere Stichprobentechniken informieren etwa Schwarz (1975), Cochran (1972, 1977) sowie Rasch et al. (1978). Innerhalb des Modells „Signifikanztest“ kommt dem Konzept der zufälligen Stichproben für die wesentlichen Ableitungen eine zentrale Bedeutung zu.

Die Verteilung der Daten in der Stichprobe wird durch die sog. „(empirische) Häufigkeitsverteilung“ angegeben. Wie die Verteilung der Rohwerte in der Population durch bestimmte Parameter gekennzeichnet ist, die die in den Daten enthaltene Information zusammenfassen und reduzieren, so sind auch die Stichprobendaten durch sog. „Stichprobenfunktionen“ oder „Statistiken“ beschreibbar, die die Häufigkeitsverteilung charakterisieren.



Statistiken stellen häufig das Stichprobenäquivalent bestimmter Parameter dar, ohne mit diesen wertmäßig übereinstimmen zu müssen. Denn durch zufallsbedingte Schwankungen weicht die Stichprobenzusammensetzung meist von der Populationszusammensetzung ab. Erhebt man nun sehr viele Stichproben der gleichen Größe  $n$  und berechnet die empirischen Werte einer Statistik, dann dokumentieren sich diese Zufallsschwankungen darin, daß die resultierende „Stichproben-Kennwerte- oder Prüf- oder Stichproben-Verteilung“ der betreffenden Statistik eine Varianz größer als Null aufweist.<sup>16)</sup>

Um die auf i. a. sehr verschiedenen Experimenten beruhenden Werte für eine Statistik problemlos auf nur eine Prüfverteilung beziehen zu können, werden die Statistiken linear transformiert, und man nennt derartige „transformierte Statistiken“ „Prüf- oder Test-Statistiken“. Häufig besteht diese Transformation in einer Standardisierung, wie es etwa beim  $t$ -Wert der Fall ist, der eine(n) standardisierte(n) Mittelwert(sdifferenz) darstellt; andere bekannte Teststatistiken sind etwa  $\chi^2$ ,  $F$  oder  $H$ .

Um von den Stichproben- auf die Populationsdaten oder -kennwerte schließen zu können, muß eine Verbindung zwischen der beobachtbaren Stichprobe und der theoretischen Grundgesamtheit konstruiert werden. Dies geschieht durch die sog. „Schätzfunktionen“, d.h. Statistiken oder Stichprobenfunktionen „mit bestimmten Eigenschaften“ - vgl. dazu Menges (1972, 294f.). Auf diese Schätzfunktionen gehen wir nicht ein; der interessierte Leser sei auf die zu Beginn dieses Teils genannte einführende Literatur verwiesen.

Wie erwähnt, weichen die Stichproben(kenn)werte in aller Regel von den Populations(kenn)werten ab - diese Abweichung oder Verzerrung wird als „unsystematisch“ (oder „zufällig“) interpretiert und als eine Folge des Stichprobenprozesses angesehen (vgl. Abschnitt 8.2). Der Signifikanztest dient nun der Beantwortung der Frage, ob ein konkretes empirisches Resultat, d.h. etwa eine empirische Realisation der Zufallsvariable „Teststatistik“, als nur durch den Zufall zustande gekommen „erklärt“ werden kann oder ob die Annahme einer systematischen „Verursachung“ angemessener erscheint. Der Erklärung „durch Zufall“ wird dann der Vorzug gegeben, wenn die Realisation der Test- oder Prüfstatistik in eine Klasse von möglichen Realisationen fällt, der eine insgesamt hohe Wahrscheinlichkeit zukommt; letztere wird unter Verwendung der o. g. Prüfverteilung bestimmt.

Für die mathematische Ableitung der Prüfverteilung muß die Gültigkeit einer statistischen Hypothese und (sehr oft auch) eine bestimmte Populationssituation (vgl. Abschnitt 7.5.2) angenommen werden. Findet man dann, daß der

<sup>16)</sup> Je größer diese Varianz der Stichprobenverteilung ist, desto geringer ist unter sonst gleichen Bedingungen die „Präzision“ eines Experiments - vgl. dazu insbesondere Abschnitt 8.4. - Mit dem sog. „Standardfehler (einer Teststatistik)“ ist die Wurzel aus dieser Varianz gemeint.

empirische Wert für eine Teststatistik in einen Bereich fällt, der unter diesen Annahmen insgesamt wenig wahrscheinlich ist, schließt man daraus, daß die als richtig unterstellte Hypothese nicht zutrifft, daß m.a. W. die untersuchte Stichprobe einer Population entstammt, deren Parameter andere Werte aufweisen bzw. deren Verteilung anders ist, als in der statistischen Hypothese angenommen.

Die statistische Hypothese, auf deren Zutreffen die Ableitung der Stichprobenverteilung einer beliebigen (Test-)Statistik beruht, heißt üblicherweise die „Null-Hypothese“ ( $H_0$ ).

Eine  $H_0$  wird in der Regel von inhaltlichen Aussagen der Form „Es besteht *kein* Unterschied zwischen verschiedenen experimentellen Behandlungen und damit deren Auswirkungen auf bestimmte Parameter bzw. Verteilungen“ impliziert.

Wir haben jedoch im Teil 6 gesehen, daß unsere Kausalhypothese  $WH_u$  zu anderen Vorhersagen führte, nämlich etwa der, daß „Unterschiede zwischen verschiedenen Kennwerten bestehen“. Man nennt die von einer derartigen inhaltlichen Aussage implizierte statistische Hypothese die „Alternativhypothese“ ( $H_1$ ).

Entschließt man sich aufgrund der experimentellen Daten, die  $H_0$  zurückzuweisen, nimmt man üblicherweise gleichzeitig die  $H_1$  an, weil die beiden statistischen Hypothesen sich gegenseitig ausschließen und darüber hinaus in der Regel so formuliert sind, daß sie alle logisch möglichen Annahmen über die Verteilungen oder Parameter ausschöpfen. Wenn wir bei unserem Beispiel aus Teil 6 bleiben, wäre die dort über Populationsmittelwerte  $\mu_k$  formulierte Hypothese  $H_1: \mu_2 - \mu_1 > 0$  um die  $H_0: \mu_2 - \mu_1 \leq 0$  zu erweitern. Entsprechend gehört zu einer  $H_0: \mu_1 - \mu_2 = 0$  die  $H_1: \mu_1 - \mu_2 \neq 0$  usf.

Man entscheidet sich also üblicherweise für die von einer WH implizierte  $H_1$ , indem man die ihr entgegengesetzte  $H_0$  zurückweist. Für diese Ablehnung der  $H_0$  wird folgendes Kriterium festgesetzt: „Weise die  $H_0$  dann zurück, wenn Du ein Ergebnis gefunden hast, das zu einer Ergebnisklasse gehört, die unter der Annahme der Gültigkeit der  $H_0$  eine Gesamtwahrscheinlichkeit aufweist, die kleiner oder gleich einem bestimmten geringen Wert  $\alpha$  ist.“ Man nennt  $\alpha$  das „Signifikanzniveau“.

Durch dieses Signifikanzniveau  $\alpha$  (auch „Umfang eines Tests“ oder Irrtumswahrscheinlichkeit“ genannt) wird die Menge aller möglichen Realisationen einer Teststatistik in zwei unterschiedlich große Teilmengen zerlegt. Die größere dieser Teilmengen heißt „Annahmereich“ und enthält diejenigen  $100(1 - \alpha)\%$  aller möglichen Werte der Prüfstatistik, die als mit der  $H_0$  vereinbar angesehen werden, deren Varianz also als zufallsbedingt angesehen werden kann. Die kleinere Teilmenge dagegen heißt „Ablehnungs- oder Rejektionsbe-

reich“, weil sie alle diejenigen Realisationen der Teststatistik enthält, die zwar unter Gültigkeit der  $H_0$  durchaus auftreten können (!), deren Auftretenswahrscheinlichkeit insgesamt jedoch so gering ist, daß sie „auf lange Sicht“ nur (sehr) selten, nämlich in  $100 \cdot \alpha$  % aller Fälle, zu erwarten sind, falls  $H_0$  zutrifft.

Man bezeichnet ein Ergebnis in diesem Ablehnungsbereich als „statistisch signifikant“ und meint damit, daß es „bedeutsam“ von einem aufgrund des Zufalls zu erwartenden Resultat abweicht.

Wenn nun eine empirische Realisation der Zufallsvariablen „Teststatistik“ in diesen Ablehnungsbereich fällt, d.h. unter Gültigkeit der  $H_0$  (sehr) unwahrscheinlich ist, dann kann dieses „seltene Ereignis“ grundsätzlich auf zwei verschiedene Arten *interpretiert* werden:

1. Die  $H_0$  ist richtig, und es ist „rein zufällig“ ein Wert aus einer Ereignisklasse mit einer insgesamt nur sehr geringen Wahrscheinlichkeit aufgetreten.
2. Das Resultat ist sehr unwahrscheinlich, falls  $H_0$  zutrifft. Es wird daher davon ausgegangen, daß  $H_0$  falsch und  $H_1$  richtig ist.

Zwischen beiden Möglichkeiten läßt sich keine objektiv richtige Entscheidung treffen; es entspricht daher lediglich einer auf Fisher (1925) zurückgehenden Konvention, wenn im Falle eines unter  $H_0$  unwahrscheinlichen (wiewohl möglichen!) Resultates entschieden wird, die  $H_0$  als falsch anzusehen - vgl. zu dieser Verfahrensweise auch Neyman (1952, 43).

An dieser Stelle wird deutlich, daß der Signifikanztest als eine spezielle Strategie aufgefaßt werden kann, mittels derer man zu Entscheidungen über das Zutreffen oder Nicht-Zutreffen von statistischen Hypothesen gelangen *kann*. Die Entscheidungskriterien sind dabei nicht „test-immanent“ oder sonstwie zwingend vorgegeben, sondern beruhen vorwiegend auf Vereinbarungen.

## 7.3 Mögliche Fehler beim statistischen Testen

### 7.3.1 Fehler unter Gültigkeit der Null-Hypothese (Fehler 1. Art)

Obwohl unter Gültigkeit der  $H_0$  extreme Resultate (sehr) unwahrscheinlich sind, besteht doch grundsätzlich die Möglichkeit ihres Auftretens. Tritt nun in empirischen Daten ein Wert der Statistik im Ablehnungsbereich unter  $H_0$  auf, obwohl  $H_0$  tatsächlich zutrifft, wird man diese irrtümlich als nicht zutreffend zurückweisen. Mit dieser (Fehl-)Entscheidung begeht man einen sog. „Fehler 1. Art“, auch  $\alpha$ -Fehler genannt.

Die bedingte Wahrscheinlichkeit für einen derartigen Fehler beträgt  $\alpha$ , denn die Wahrscheinlichkeit der Klasse von sehr extremen Resultaten, die im Ab-

lehnungsbereich von  $H_0$  liegen, ist bei stetigen Verteilungen genau und bei diskreten Verteilungen höchstens gleich dem Signifikanzniveau  $\alpha$ :

$$p(\text{Zurückweisung von } H_0 \mid H_0 \text{ trifft zu}) = \alpha.$$

Diese Fehlerwahrscheinlichkeit kann durch die Wahl sehr kleiner Werte für  $\alpha$ , etwa  $\alpha = 0,05$  oder  $\alpha = 0,01$  etc., vom E selbst kontrolliert und gering gehalten werden. Mit der Fixierung des numerischen Wertes für  $\alpha$  expliziert der E seine maximale Bereitschaft, einen Fehler 1. Art zu begehen. Mit der Wahl des Wertes für  $\alpha$  wird der Annahme- und Ablehnungsbereich für die betr. Stichprobenverteilung ebenfalls festgelegt.

Diese Festlegung hat in jedem Fall vor der Datenerhebung zu erfolgen, da durch sie der Signifikanztest „erst . . . eindeutig bestimmt ist“ (Haagen & Seifert, 1979, 203).

Das Komplement zur Wahrscheinlichkeit  $\alpha$  ist  $1 - \alpha$ ; dabei handelt es sich um die bedingte Wahrscheinlichkeit, eine zutreffende  $H_0$  auch als richtig auszuweisen:

$$p(\text{Annahme von } H_0 \mid H_0 \text{ trifft zu}) = 1 - \alpha.$$

Die Gesamtwahrscheinlichkeit aller numerischen Werte der Statistik im Annahmebereich unter  $H_0$  beträgt  $1 - \alpha$ ; d.h. in  $100 \cdot (1 - \alpha) \%$  aller Fälle wird erwartet, daß der Signifikanztest eine in der Population wahre Null-Hypothese „entdeckt“.

Nun ist hier fortwährend von „der“ Null-Hypothese die Rede, obwohl wir im vorangegangenen Abschnitt 7.2 gesehen haben, daß unter „der“ Null-Hypothese auch eine ganze Klasse von Werten für den in Frage stehenden Parameter spezifiziert werden kann; dies kommt etwa in den folgenden (zulässigen) Hypothesenformulierungen zum Ausdruck:

$$H_0: \mu_2 - \mu_1 \leq 0 \text{ oder: } H_0: \mu \leq \mu_0 = 100 \text{ (siehe vorn).}$$

Man nennt derartige Hypothesen, die anstelle nur eines einzigen Wertes einen ganzen Wertebereich für den Parameter spezifizieren, „zusammengesetzt“ (auch „unspezifisch“ und „unexakt“). Dagegen heißt eine Hypothese, die nur genau einen Wert für den Parameter zuläßt, „einfach“ (auch „exakt“ und „spezifisch“). Die folgenden beiden Hypothesen sind bspw. exakt:

$$H_0: \mu_2 - \mu_1 = 0 \text{ oder: } H_0: \mu = \mu_0 = 100.$$

Näheres zu diesen Unterscheidungen findet man etwa bei Stilson (1966, 385f.), Weber (1967, 167f.) und Bortz (1979, 148f.); zu beachten ist hierbei, daß der Sprachgebrauch recht uneinheitlich ist, wie die Arbeiten von Hays (1977, 335f.) und Haagen & Seifert (1979, 206-225) belegen.

Die Frage stellt sich also, welche von diesen theoretisch beliebig vielen Null-Hypothesen dem statistischen Test unterzogen wird. Grundsätzlich wird auch dann, wenn  $H_0$  eine zusammengesetzte Hypothese ist, nur eine einzige davon statistisch geprüft, und zwar diejenige einfache  $H_0$ , die dem Parameterbereich am nächsten liegt, der von allen unter der Bezeichnung „ $H_1$ “ zusammengefaßten Hypothesen abgedeckt wird - vgl. dazu im einzelnen Neyman & Pearson (1933 a, b) und Haagen & Seifert (1979, 206-225).

In den o. gen. Beispielen werden also nur die folgenden Hypothesen geprüft:

$$H_0: \mu_2 - \mu_1 = 0 \text{ und } H_0: \mu = \mu_0 = 100.$$

Alle anderen unter der Klasse „ $H_0$ “ zusammengefaßten Hypothesen sind dann zu verwerfen, wenn auch die einfache Hypothese zurückgewiesen werden kann, die den relativ extremsten Wert für den interessierenden Parameter angibt.

Die festgelegte Wahrscheinlichkeit für einen Fehler 1. Art (a-Fehler) ist dabei maximal für die getestete einfache  $H_0$  und geringer für alle anderen Elemente der Klasse „ $H_0$ “ (Bredenkamp, 1972, 21 f.; Hays, 1977, 363).<sup>17)</sup>

Die vorstehenden Ausführungen über den a-Fehler beruhen auf der Voraussetzung, daß in der Population eine einfache statistische Null-Hypothese zutrifft. Bei den folgenden Darstellungen wollen wir davon ausgehen, daß diese Annahme nicht richtig ist, sondern daß die unter der Alternativhypothese formulierten Annahmen der Populationssituation entsprechen.

### 7.3.2 Fehler unter Gültigkeit der Alternativhypothese (Fehler 2. Art)

Die statistische Alternativhypothese  $H_1$  wird in der Regel von einer wissenschaftlichen Hypothese wie unserer Beispielhypothese  $WH_u$  impliziert, die wir u.a. auf den Seiten 68 und 69 bereits behandelt haben.

Mit der  $H_1$  werden Annahmen über die Werte des interessierenden Parameters spezifiziert, die grundsätzlich von den unter  $H_0$  formulierten abweichen. Trifft daher eine der unter der Klasse „ $H_1$ “ angegebenen einfachen Alternativhypothesen zu, resultieren (fast) stets Stichprobenverteilungen der benutzten Prüfstatistik, die von den unter  $H_0$  geltenden Verteilungen quantitativ abweichen.

Der Grad dieser Abweichung von den tabellierten und zur Signifikanzbeurteilung benutzten sog. „zentralen“ Prüfverteilungen kann meist durch einen sog.

---

<sup>17)</sup> Wenn im folgenden weiter von der Null-Hypothese die Rede ist, soll darunter ggf. die gesamte Hypothesenklasse „ $H_0$ “ verstanden werden, ohne daß explizit zwischen „einfachen“ und „zusammengesetzten“ Hypothesen unterschieden wird.

„Nicht-Zentralitätsparameter“ angegeben werden (vgl. dazu im einzelnen Abschnitt 9). Man bezeichnet die unter Gültigkeit der  $H_1$  resultierenden Stichprobenverteilungen als die „nicht-zentralen Verteilungen“ der betreffenden Teststatistik, die für  $\chi^2$  und F erstmals Fisher (1928) näher erörtert hat.

Auch unter Gültigkeit der  $H_1$  sind die Realisationen der Teststatistik bestimmten Zufallsschwankungen unterworfen, die im konkreten Einzelfall so ausgeprägt sein können, daß der Wert der Statistik im Annahmehereich unter  $H_0$  liegt, die dann angenommen wird.

Da aber nach unserer Voraussetzung  $H_1$  richtig ist, trifft der E damit eine falsche Entscheidung, er begeht einen „Fehler 2. Art“. Die Wahrscheinlichkeit für einen derartigen Fehler 2. Art, d.h. für eine irrtümliche Entscheidung für  $H_0$ , wird formal mit  $\beta$  bezeichnet; entsprechend findet sich daher auch oft die Bezeichnung „ $\beta$ -Fehler“.

Das Komplement zu  $\beta$ , also  $1 - \beta$ , heißt „Teststärke“ („Macht“, „Mächtigkeit“, „Güte“, „Trennschärfe“ oder „Power“) eines Tests und gibt die Wahrscheinlichkeit einer richtigen Entscheidung für  $H_1$  an:

$$\begin{aligned} p(\text{Zurückweisung der } H_1 \text{ und Annahme der } H_0 \mid H_1 \text{ trifft zu}) &= \beta \\ p(\text{Annahme der } H_1 \mid H_1 \text{ , trifft zu}) &= 1 - \beta \end{aligned}$$

Wir haben im vorigen Abschnitt festgestellt, daß mit jeder der unter „ $H_0$ “ zusammengefaßten einfachen Hypothesen eine „eigene“ (Gesamt-)Irrtumswahrscheinlichkeit  $\alpha$  verbunden ist, deren Maximalwert üblicherweise vom E durch eine Festsetzung gering gehalten wird. In der gleichen Weise ist auch mit jeder der unter „ $H_1$ “ zusammengefaßten einfachen Hypothesen eine eigene Fehlerwahrscheinlichkeit  $\beta$  verknüpft (vgl. zu Einzelheiten etwa Menges, 1972, 328-335, oder Hays, 1977, 357-373). Die Wahrscheinlichkeit  $\beta$  für eine irrtümliche Annahme der  $H_0$  läßt sich daher nur dann bestimmen, wenn angegeben werden kann, welche der zahlreichen einfachen Hypothesen unter „ $H_1$ “ zutrifft, wenn also m.a. W. bekannt ist, wie sehr der interessierende Parameter (resp. die Verteilung) von dem unter  $H_0$  spezifizierten Wert abweicht. Wir bezeichnen diese Abweichungen im folgenden allgemein als „experimentelle Effekte“ (EE), deren Auftreten trotz dieser Benennung nicht an die Art der Untersuchung gebunden ist. Auf diese experimentellen Effekte gehen wir im Teil 9 ausführlich ein.

Bestimmt man unter Verwendung der nicht-zentralen Verteilungen die Werte für die Teststärke  $1 - \beta$  in Abhängigkeit vom Signifikanzniveau  $\alpha$ , dem experimentellen Effekt und der Stichprobengröße  $N$  für einen bestimmten Test, erhält man die sog. „Teststärke-“, „Güte-“ oder „Trennschärfefunktion“ des betr. Signifikanztests - zu Einzelheiten siehe etwa Menges (a. a. O.) und Hays (a. a. O.).

Diese Gütefunktionen sind für einige der gebräuchlichen (sog. „parametrischen“) Teststatistiken wie  $\chi^2$ , F und t seit geraumer Zeit in tabellarischer Form als „Power Tables“ sowie in graphischer Form als „Power Charts“ verfügbar, allerdings vornehmlich in mathematischen und statistischen Fachzeitschriften (vgl. Abschnitt 10.2 und 10.3). Entsprechend wurden sie von Psychologen kaum rezipiert; und auch in den meisten Lehrbüchern der Statistik für Psychologen und Sozialwissenschaftler wird das Konzept „Teststärke“ - wenn überhaupt - nur kurz behandelt.“) Es kann daher kaum verwundern, wenn in der üblichen Forschungspraxis die Wahrscheinlichkeit für einen Fehler 2. Art ( $\beta$ ) nicht kontrolliert oder bestimmt wird.

Diese Unterlassung kann bei der Entscheidung zwischen den statistischen Hypothesen zu schwerwiegenden Konsequenzen führen.

1. Möglichkeit: Der E entscheidet sich aufgrund des Tests für  $H_0$ . Diese Entscheidung kann deshalb zustande gekommen sein, weil  $H_0$  tatsächlich zutrifft, oder aber weil zwar die  $H_1$  richtig ist, dem E jedoch ein Fehler 2. Art unterlaufen ist. Da die Wahrscheinlichkeit  $\beta$  für diesen Fehler meist nicht bekannt ist, kann auch nichts über die Güte der Entscheidung ausgesagt werden. Zwar beträgt unter der Annahme der Richtigkeit der  $H_0$  die Wahrscheinlichkeit eines Resultates im Annahmehereich  $1 - \alpha$ , aber unter der Richtigkeit der  $H_1$  kann die Wahrscheinlichkeit, ein Ergebnis im Ablehnungsbereich von  $H_1$  zu erhalten, also  $\beta$ , grundsätzlich numerisch genau so groß wie die Wahrscheinlichkeit  $1 - \alpha$  werden.
2. Möglichkeit: Der E lehnt  $H_0$  ab und entscheidet sich für die  $H_1$ . Diese Entscheidung nun kann deswegen getroffen worden sein, weil in der Tat  $H_1$  zutrifft - in diesem Fall ist die Wahrscheinlichkeit eines Resultates im Annahmehereich von  $H_1$  gleich  $1 - \beta$  -, oder aber, weil dem HE ein Fehler 1. Art unterlaufen ist. Zwar wird die Wahrscheinlichkeit dieses Fehlers vom E durch die Wahl eines kleinen Signifikanzniveaus gering gehalten; da aber andererseits die Wahrscheinlichkeit  $1 - \beta$ , eine zutreffende  $H_1$  als richtig auszuweisen, nicht bekannt ist, kann prinzipiell nicht ausgeschlossen werden, daß ihre numerische Größe gleich der oder kleiner als die Wahrscheinlichkeit  $\alpha$  ist.

Wenn wir uns die Implikationsbeziehung zwischen der wissenschaftlichen Hypothese  $WH_u$  und der statistischen Hypothese  $H_1$ , also  $WH_u \rightarrow H_1$ , noch einmal vergegenwärtigen, ergeben sich aus den vorstehenden Erörterungen bestimmte (mögliche) Konsequenzen für die Beurteilung der  $WH_u$ :

Begehen wir einen Fehler 1. Art, besteht die Gefahr, daß die wissenschaftliche Hypothese fälschlicherweise als bewährt angesehen wird. Begehen wir dage-

<sup>18)</sup> Der interessierte Leser mag selbst Belege für diese Behauptung beibringen; dies stellt jedoch erfahrungsgemäß keine Schwierigkeit dar.

gen einen Fehler 2. Art, besteht die Gefahr, daß die wissenschaftliche Hypothese irrtümlich falsifiziert wird.“)

Wie bereits im Abschnitt 6 angesprochen, ist es unabdingbar, eine zutreffende statistische Hypothese mit großer Wahrscheinlichkeit annehmen und eine nicht-zutreffende mit großer Wahrscheinlichkeit ablehnen zu können, da nur auf diese Weise - langfristig - gerechtfertigte Falsifikationen oder aber Bestätigungen von wissenschaftlichen Hypothesen zu erwarten sind.

Es ist daher zwingend erforderlich, nicht nur die Wahrscheinlichkeit  $\alpha$  für einen Fehler 1. Art zu kontrollieren und gering zu halten - hierdurch wird der Gefahr fälschlicher Bestätigungen der  $WH_u$  entgegengewirkt -, sondern auch die Wahrscheinlichkeit  $\beta$  für einen Fehler 2. Art zu kontrollieren und gering zu halten - hierdurch wird die Gefahr ungerechtfertigter Falsifikationen in Grenzen gehalten.“)

Diese Forderung nach Kontrolle *beider* Fehlerwahrscheinlichkeiten sind nicht neu; sie wurde bereits von Sterling (1960), Smart (1964), Bakan (1966), Brendenkamp (1969b, 1972, 1975, 1979, 1980), Krause & Metzler (1978) und Witte (1977, 1980) erhoben, um nur einige Autoren zu nennen. Die aus ihr folgenden konkreten Handlungsanweisungen für die Praxis werden allerdings selten realisiert. Bevor wir uns mit diesen näher befassen, wollen wir zunächst die Größen zusammenstellen, die den Ausgang eines (beliebigen) Signifikanztests determinieren, und anschließend einen Blick auf die verbreitete Forschungspraxis werfen.

## 7.4 Die Determinanten eines Signifikanztests

Unseren bisherigen Darstellungen kann entnommen werden, daß der Ausgang eines Signifikanztests stets vornehmlich von den folgenden vier Größen abhängig ist:

1. dem Signifikanzniveau  $\alpha$ ,
2. der Teststärke  $1 - \beta$ ,
3. dem in den empirischen Daten enthaltenen experimentellen Effekt  $EE$ ,
4. der Varianz der Stichprobenverteilung der gewählten Teststatistik, die ihrerseits stets von der Stichprobengröße  $N$  resp.  $n$  abhängt (vgl. zu Einzelheiten Abschnitt 8.3).

Die Beziehungen zwischen diesen vier Größen werden üblicherweise in den Grundlagen-Texten nur am Rande behandelt (vgl. jedoch u.a. Hays, 1963,

---

<sup>19)</sup> Selbstverständlich treffen diese Ausführungen mutatis mutandis auch in den Fällen zu, in denen die wissenschaftliche Hypothese eine statistische Null-Hypothese impliziert:  $WH_z \rightarrow H_0$  (Hager & Westermann, im Druck, b).



1977; Leiser, 1978; Haagen & Seifert, 1979) und in der Praxis (zu) wenig beachtet.

Um sich eine zumindest ungefähre Vorstellung über die Bedeutung zu verschaffen, der insbesondere der Stichprobengröße  $N$  für den Ausgang eines Signifikanztests, d.h. der Beurteilung eines empirischen Resultates auf statistische Signifikanz, zukommt, sollte der Leser sich mit den entsprechenden Beispielen bei Leiser (1978, 178-191) oder auch bei Hays (1977, e.g. 357-362) vertraut machen. Zur Verdeutlichung der für uns wichtigsten Beziehungen wollen wir uns kurz mit den Tabellen C 4 und C 5 in Gaensslen & Schubö (1973, 1976, 314-317) befassen (vgl. ersatzweise etwa Fisher & Yates, 1963, 63, oder Bortz, 1979, 832f.). Mittels dieser Tabellen kann der empirisch erhaltene Wert für einen Produkt-Moment-Korrelationskoeffizienten auf statistische Signifikanz beurteilt werden.

Aus den Tabellen wählen wir einige beliebige Werte für  $r$  aus, stellen fest, ob diese Werte bei vorgegebener Stichprobengröße  $N$  und vorgegebenem Signifikanzniveau  $\alpha$  als statistisch signifikant (s.) oder nicht signifikant (n.s.) beurteilt werden. Diese Informationen sind in den Tabellen 7.1 und 7.2 neu zusammengestellt und um eine Angabe über den jeweiligen experimentellen Effekt  $EE$  erweitert worden, der hier gleich dem Quadrat des Korrelationskoeffizienten ist.

Tabelle 7.1: Statistische Signifikanz von  $r$  bei  $\alpha = 0.05$ .

$r$	$EE = r^2$	Stichprobengröße $N$					
		5	8	16	32	400	1000
0,90	0,81	S.	S.	S.	S.	S.	S.
0,35	0,12	n.s.	n.s.	n.s.	S.	S.	S.
0,12	0,01	n.s.	n.s.	n.s.	n.s.	S.	S.
0,08	0,006	n.s.	n.s.	n.s.	n.s.	n.s.	S.

Tabelle 7.2: Statistische Signifikanz von  $r$  bei  $\alpha = 0,01$ .

$r$	$EE = r^2$	Stichprobengröße $N$					
		5	8	16	32	400	1000
0,90	0,81	n.s.	S.	S.	S.	S.	S.
0,35	0,12	n.s.	n.s.	n.s.	n.s.	S.	S.
0,12	0,01	n.s.	n.s.	n.s.	n.s.	n.s.	S.
0,08	0,006	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Die Schlüsse, die aus den in den o.a. Tabellen enthaltenen Informationen zu ziehen sind, können ungeachtet der jeweils benutzten Teststatistik Gültigkeit für alle Signifikanztests beanspruchen:

1. Ist ein Ergebnis statistisch signifikant geworden, bedeutet dies lediglich, daß ein überzufälliger experimenteller Effekt im Experiment aufgetreten ist.<sup>20)</sup> Wie groß dieser ist, kann durch die statistische Signifikanz nicht festgestellt werden.
2. Jeder beliebig kleine EE kann ungeachtet seiner inhaltlichen Bedeutung statistisch „signifikant gemacht“ werden (Kleiter, 1969, 150f.), indem man eine genügend große Stichprobe benutzt. Unter sonst gleichen Bedingungen ist die statistische Signifikanz ausschließlich von der Stichprobengröße abhängig. Die Beispiele, die etwa Nunnally (1960), Bakan (1966), Meehl (1967), Kleiter (1969) und Cowles (1974) angeben, belegen, daß diese Beziehung in der Praxis offenbar sehr wohl bekannt ist und eingesetzt wird, weil aus noch anzuspreekenden Gründen signifikante Resultate die einzig erwünschten Ausgänge von Signifikanztests darstellen (vgl. hierzu Abschnitt 7.4.1).
3. Je kleiner ein EE ist, desto größer muß der Stichprobenumfang  $N$  sein, um diesen EE als statistisch signifikant ausweisen zu können.
4. Je kleiner ein EE ist, desto geringer ist unter sonst gleichen Bedingungen die Teststärke  $1 - \beta$ , wie zwar nicht den o.a. Tabellen, wohl aber den Gütefunktionen zu entnehmen ist. Eine Erhöhung der Teststärke ist dann über die Erhöhung des Signifikanzniveaus und/oder des Stichprobenumfanges möglich oder aber durch die Wahl eines anderen Versuchsplanes und/oder Auswerteverfahrens.

Die weiteren Beziehungen mag sich der interessierte Leser selbst ableiten. Wir wollen uns statt dessen im folgenden kurz der Frage zuwenden, warum der statistischen Signifikanz gemeinhin eine überragende Bedeutung beigemessen wird.

#### 7.4.1 Forschungs- und Publikationspraxis 1: Signifikanzniveau und $p$ -Werte

Bei Durchsicht psychologischer Fachzeitschriften wird man sehr leicht feststellen können, daß bei der Darstellung von empirischen Resultaten meist weder von einer Irrtumswahrscheinlichkeit  $\alpha$  noch von der Möglichkeit eines Fehlers 2. Art die Rede ist. Statt dessen findet man sog. „ $p$ -Werte“, die die Wahrscheinlichkeit des gefundenen und aller (noch) weiter von  $H_0$  abweichenden Resultate unter der Voraussetzung angeben, daß die  $H_0$  zutrifft. Diese  $p$ -Werte werden nach der Berechnung des empirischen

---

<sup>20)</sup> Das Auftreten eines EE ist nicht an die statistische Signifikanz gebunden. Wie Hays (1963, 326) im einzelnen darlegt und begründet, stellt das völlige Fehlen eines EE in der Empirie eine Ausnahme dar.

Wertes der benutzten Teststatistik entsprechenden Tabellen (etwa Pearson & Hartley, 1954, 1972; Fisher & Yates, 1963) entnommen und entsprechen im Grunde einem „gleitenden Signifikanzniveau“. Sie werden dann ungeachtet geringfügiger Unterschiede bei einzelnen Autoren i. a. nach folgendem Schema „beurteilt“ - vgl. dazu Harnatt (1975, 603), Krause & Metzler (1978, 225), Bortz (1979, 146) und Haagen & Seifert (1979, 203):

- $p = 0,1$  „bedeutet“: „symptomatische Abweichung“;<sup>21)</sup>  
 $p = 0,05$  „bedeutet“: „signifikante Abweichung“, oft nur durch einen Stern („\*“) gekennzeichnet;  
 $p = 0,01$  „bedeutet“: „sehr“ oder „hoch signifikante Abweichung“, meist durch zwei Sterne („\*\*“) symbolisiert;  
 $p = 0,001$  „bedeutet“: „äußerst“ oder „höchst signifikante Abweichung“, abgekürzt durch drei Sterne („\*\*\*“) angegeben.

Wie Rosenthal & Gaito (1963, 1964) sowie Beauchamp & May (1964) experimentell glaubhaft machen konnten, ist das (subjektive) Vertrauen in die Ergebnisse von Signifikanztests um so größer, je kleiner die nachträglich ermittelten  $p$ -Werte sind.

Diese Interpretation der Wahrscheinlichkeiten bestimmter extremer Ereignisklassen unter Gültigkeit der statistischen Null-Hypothese ist dabei nicht ausschließlich charakteristisch für die untersuchten „graduate students“ und „Ph.D.'s“ (Rosenthal & Gaito, 1963, 33; Beauchamp & May, 1964, 272), sondern auch für Herausgeber von Zeitschriften. Offenbar ist nämlich die Chance, daß ein Artikel, der eine statistische Auswertung via Signifikanztest enthält, zur Publikation akzeptiert wird, um so größer, je kleiner der ermittelte  $p$ -Wert ist - man vergleiche hierzu insbesondere die durch die einflußreiche Arbeit von Bakan (1966) bekannt gewordene sog. „Herausgeber-Philosophie“ von A. Melton (1962), der im Zeitraum von 1951-1962 das „Journal of Experimental Psychology“ herausgab.

Melton stellt dabei keine Ausnahmeerscheinung, sondern eher einen „typischen“ Vertreter für die weitverbreitete Fehlinterpretation des nachträglich bestimmten  $p$ -Wertes dar, wie u.a. durch die (kritischen) Publikationen von Sterling (1959), Cohen (1962, 1965), Smart (1964), Bakan (1966), Skipper, Guenther & Nass (1967), Labovitz (1968), Bozarth & Roberts (1972) sowie Schulman, Kupst & Suran (1976) im einzelnen dokumentiert wird.

Auf die nicht zu unterschätzenden Folgen dieser offenbar unausrottbaren Fehlinterpretation der statistischen Fehlerwahrscheinlichkeit  $\alpha$  beim Hypothesentesten als Maß für die Größe oder inhaltliche Bedeutsamkeit eines experimentellen Effektes haben neben den o. gen. Autoren insbesondere Bredenkamp (1972, 51-73), Greenwald (1975), Lane & Dunlap (1978) sowie Rosenthal (1979) eindringlich hingewiesen.

Besonders bemerkenswert ist bei diesem Phänomen allerdings, daß diesen speziellen Interpretationen durch die Darstellungen in vielen Einführungslehrbüchern zur Statistik und/oder von renommierten Autoren entschieden Vorschub geleistet wird; Beispie-

---

<sup>21)</sup> Anstelle des in der Literatur gebräuchlichen Ausdrucks „Abweichung“ benutzen wir in dieser Arbeit die Bezeichnung „experimenteller Effekt“.

le für diese Behauptung finden sich u.a. in Bredenkamp (1972, 57), Harnatt (1975, 600f.) sowie Haagen & Seifert (1979, 202f., 243); siehe ferner auch Guttman (1977, 91 f.) und Bortz (1979, e.g. 143).

#### 7.4.2 Forschungs- und Publikationspraxis II: Experimentelle Effekte und Teststärke

Aufgrund der Ausführungen im vorangegangenen Abschnitt und der unzureichenden Beachtung, die der Teststärke (und den experimentellen Effekten) in der überwiegenden Mehrzahl der Lehrbücher gewidmet wird<sup>22)</sup> nimmt es nicht wunder, daß üblicherweise Angaben über die Teststärke (und den EE) nicht publiziert werden, obwohl dies für die EE von einigen Zeitschriften-Herausgebern bereits gefordert wird (vgl. Bredenkamp & Feger, 1970; ferner Lane & Dunlap, 1978; Soderquist & Hussian, 1978).

Allerdings läßt sich die Teststärke im Rahmen von statistischen Reanalysen für zahlreiche experimentelle Anordnungen und Auswerteverfahren auch nach der Datenerhebung und -auswertung noch bestimmen, weil sie bei vorgegebenem Signifikanzniveau  $\alpha$ , fester Stichprobengröße  $N$  und bekannter Stichprobenverteilung ausschließlich von der Größe des EE in den erhobenen Daten abhängt (s.o.; zur Bestimmung des EE aus den Daten siehe Abschnitt 9.3.2.3 und 9.4). Diese Tatsache haben sich zahlreiche Autoren zunutze gemacht und die Teststärke für publizierte Daten und Experimente analysiert, die in verschiedenen Bereichen der psychologischen Forschung und einigen verwandten Gebieten wie etwa der Kommunikationsforschung durchgeführt worden sind - vgl. u.a. Cohen (1962, 1965, 1973 b), Brewer (1972), Katzer & Sodt (1973), Chase & Tucker (1975), Kroll & Chase (1975), Chase & Baran (1976), Chase & Chase (1976), Schmidt, Hunter & Urry (1976), Treiber (1977), Treinies (1977) sowie Cascio, Valenzi & Silbey (1978, 1980). Diesen Arbeiten läßt sich übereinstimmend entnehmen, daß in der Mehrzahl der reanalysierten Untersuchungen die Teststärke  $1 - \beta$  (sehr) gering war.

Als weiteres Resultat der genannten Reanalysen ist bemerkenswert, daß die in den Experimenten aufgetretenen experimentellen Effekte fast durchgängig nur „klein“ gewesen sind - ein kleiner EE entspricht etwa einem Wert von 0,01 in den Tabellen 7.1 und 7.2 (vgl. zu näheren Einzelheiten die Abschnitte 10.3.3.2

---

<sup>22)</sup> Einer der Gründe hierfür mag darin zu suchen sein, daß derjenige der bedeutenden Statistiker, der die heutzutage praktizierte Form der statistischen Inferenz wohl am stärksten geprägt hat, R. A. Fisher, dieses Konzept strikt ablehnte. Er schrieb dazu in seinem Werk „The design of experiments“ (1935, 1951, 17): „The notion of an error of the so-called ‚second kind‘, due to accepting the null hypothesis »when it is false« . . . has no meaning with respect to simple tests of significance, in which the only available expectations are those which flow from the null hypothesis being true.“

und 11.1.3). Wegen der erwähnten Beziehung zwischen der Teststärke und dem EE (s.o. und Abschnitt 7.4) ist dieser Befund nicht verwunderlich. Angemerkt sei noch, daß die publizierten EE fast ausnahmslos als statistisch signifikant beurteilt worden waren.

Solange die Größe eines aufgedeckten experimentellen Effektes unter ausschließlicher Verwendung des (nachträglich bestimmten) p-Wertes „beurteilt“ wird, kann die Bedeutung der einzelnen empirischen Befunde für die zu prüfende wissenschaftliche Hypothese nicht abgeschätzt werden, und vor diesem Hintergrund muß der folgenden Aussage von Guttman (1977, 92) im Prinzip zugestimmt werden: „No one has yet published a scientific law in the social sciences which was developed, sharpened, or effectively substantiated on the basis of tests of significance.“

#### *7.4.3 Forschungs- und Publikationspraxis III: Entwicklung einer vorläufigen Zielvorstellung*

Aus unseren bisherigen Ausführungen geht u.a. hervor, daß von den vier Determinanten eines jeden Signifikanztests sowohl das Signifikanzniveau  $\alpha$  als auch die Teststärke  $1 - \beta$  vom E selbst zu kontrollieren und damit vor einem Experiment festzulegen sind; m.a.W. muß der E jeweils angeben, wie groß seine maximale Bereitschaft ist, einen Fehler 1. und einen Fehler 2. Art zu begehen.

Folgt der E diesem Vorgehen, ist der Ausgang des Signifikanztests nur noch von der tatsächlichen Größe des EE und dem Stichprobenumfang  $N$  abhängig.

Nun ist der Stichprobenumfang  $N$  in der Praxis meist eine Größe, die sich nach der Anzahl der jeweils (ad hoc) zur Verfügung stehenden Untersuchungseinheiten ( $V_{pn}$ ) ergibt (siehe u.a. Smart, 1966; Higbee & Wells, 1972; Oakes, 1972). Aus dieser Tatsache ergeben sich die folgenden Probleme (vgl. auch Teil 4):

- (1) Wenn die Stichprobengröße i.a. nach dem „Kriterium“ der Bereitwilligkeit und Verfügbarkeit der  $V_{pn}$  fixiert wird, ist die Größe des bei vorgegebenen Höchstwerten für  $\alpha$  und  $\beta$  im Experiment entdeckbaren EE vom E nicht kontrollierbar. Es sollte jedoch deutlich geworden sein, daß es gerade die *Größe* des EE ist, der die besondere Aufmerksamkeit des E gelten muß, denn diese Größe stellt einen wesentlichen Bestandteil der Interpretation der experimentellen Daten im Hinblick auf die wissenschaftliche Hypothese dar (vgl. zu Einzelheiten insbesondere Teil 11).

Man kann sich diesen Zusammenhang auf einer intuitiven Basis verdeutlichen, indem man sich überlegt, daß eine  $WH_{ij}$ , die für eine spezifische Versuchsanordnung einen EE

in der Größenordnung  $r^2 = 0,60$  vorhersagt<sup>23)</sup>, wesentlich „attraktiver“ ist als eine konkurrierende Hypothese  $WH_v$  zur Erklärung des gleichen Phänomens, die aber nur einen EE der Größe  $r^2 = 0,10$  erwarten läßt. Die  $WH_u$  ist attraktiver, weil sie größere Anteile der Datenvarianz „aufzuklären“ vermag und damit auch genauere Vorhersagen ermöglicht als die konkurrierende  $WH_v$ . Dieser Aspekt ist sowohl im Hinblick auf das Ziel einer möglichst genauen und sparsamen wissenschaftlichen Erklärung empirischer Sachverhalte von Bedeutung als auch für die Ableitung von technologischen Prognosen (siehe dazu Teil 5) und würde bei ausschließlicher Betrachtung der statistischen Signifikanz weitestgehend unbeachtet bleiben. Anzumerken ist hierbei noch, daß dem EE auch dann eine zentrale Bedeutung zukommt, wenn die psychologische Hypothese keine quantitativen Prognosen zuläßt, weil auch in diesem Fall festgestellt werden kann, wieviel Datenvarianz sie im Vergleich zu konkurrierenden Hypothesen „aufklärt“ - siehe dazu im einzelnen Teil 9.

Diese Ausführungen legen insgesamt nahe, zusätzlich zu den maximalen Fehlerwahrscheinlichkeiten  $\alpha$  und  $\beta$  als weiteres Kriterium auch noch den erwarteten experimentellen Effekt als Mindesteffekt (EEM) festzulegen und unter Benutzung der Teststärkefunktionen die Stichprobengröße so zu bestimmen, daß dieser EEM auch mit der vorgegebenen Teststärke im Experiment entdeckt wird, sofern er tatsächlich vorhanden ist. Im Teil 10 werden wir Strategien vorstellen, die die Bestimmung des Stichprobenumfanges nach diesen Kriterien ermöglichen, und im Teil 11 werden wir darauf eingehen, auf welche Arten man zu Festlegungen des EEM gelangen kann.

Wenden wir uns zunächst jedoch dem zweiten Problem zu, das sich aus bestimmten üblichen Praktiken ergeben kann:

- (2) Die zur Verfügung stehenden  $V_{pn}$  liefern in der Regel keine Daten, die man als zufällige Stichprobe im oben definierten Sinne auffassen kann. Welche Rechtfertigung gibt es angesichts dieser Tatsache für den Einsatz von Signifikanztests?

Obwohl wir diese Frage erst im Abschnitt 8.2.6 behandeln werden, wollen wir im folgenden Abschnitt 7.5 einige zu ihrer Beantwortung relevante Vorinformationen zusammentragen, indem wir u.a. auf Signifikanztests eingehen, deren valider Einsatz nicht an die Voraussetzung von Zufallsstichproben gebunden ist. Des weiteren sprechen wir dort Tests an, deren Anwendung insgesamt an schwächere Voraussetzungen gebunden ist, als dies bei den bisher angesprochenen Verfahren (wie  $t$ -,  $\chi^2$ - und  $F$ -Test) der Fall ist; die Kenntnis bestimmter Eigentümlichkeiten dieser sog. „nicht-parametrischen und verteilungsfreien“ Verfahren ist für verschiedene nachfolgende Ausführungen von besonderem Interesse.

---

<sup>23)</sup> Quadrierte Korrelationen ( $r^2$ ) sind, wie wir im Teil 9 noch sehen werden, spezielle Maße für experimentelle Effekte (vgl. auch Abschnitt 7.4).

## 7.5 Arten statistischer Hypothesen und ihre Prüfung

Neben den bereits eingeführten Differenzierungen zwischen einer statistischen Null- und einer Alternativhypothese (siehe Abschnitt 7.2 und 7.3) sowie zwischen einfachen und zusammengesetzten Hypothesen (Abschnitt 7.3.1) sind weitere Unterscheidungen zu beachten, auf die wir im folgenden eingehen wollen.

### 7.5.1 Gerichtete und ungerichtete Hypothesen und ihre Prüfung

Eine „gerichtete statistische Hypothese“ legt eine Ordnung für die zu vergleichenden Parameter fest; sie spezifiziert die „Richtung“ eines Unterschiedes etwa wie folgt:

$$H_1: \mu_1 < \mu_2 \text{ oder } H_1: \mu_1 - \mu_2 < 0.$$

Dagegen sagt eine „ungerichtete statistische Hypothese“ nur etwas über Unterschiede zwischen Parametern in beliebiger Richtung aus, etwa in dieser Form:

$$H_1: \mu_1 \neq \mu_2 \text{ oder } H_1: \mu_1 - \mu_2 \neq 0.$$

Diesen Hypothesenarten entsprechen unterschiedliche wissenschaftliche Hypothesen; auf die Implikationsbeziehungen gehen wir im Abschnitt 8.1 näher ein.

Von der Art und Weise, wie eine statistische Hypothese formuliert wird, muß unterschieden werden, wie sie aufgrund der zur Verfügung stehenden Stichprobenverteilungen geprüft wird, nämlich entweder „einseitig“ oder „zweiseitig“.

Diese Termini beziehen sich darauf, ob an beiden „Enden“ („Seiten“) einer Stichprobenverteilung je ein Rejektionsbereich definiert wird (zweiseitiger Test) oder nur an einem „Ende“ (einseitiger Test) (vgl. u.a. Edwards, 1971, 109-112; Hays, 1977, 369-374; Bortz, 1979, 150-152).

Ein einseitiger Test führt dabei unter sonst gleichen Bedingungen eher zur Ablehnung der geprüften  $H_0$  als ein zweiseitiger Test. Allerdings ist nur eine begrenzte Anzahl von statistischen Verfahren verfügbar, die dem Experimentator die Wahl zwischen einem oder zwei Ablehnungsbereichen ermöglichen; zu diesen zählen etwa die Binomial-Tests und der t-Test. Besteht die Wahlmöglichkeit bzgl. des Rejektionsbereiches, werden gerichtete statistische Hypothesen adäquat unter Verwendung einseitiger Tests geprüft - vgl. Abschnitt 8.1.2. In manchen Fällen bestimmt (leider) das gewählte Testverfahren, ob der E einen oder zwei Rejektionsbereiche festlegen kann.

So eignet sich die F-Verteilung zwar zur Beurteilung von gerichteten Hypothesen über Varianzen (vgl. Hays, 1977, 445-447); dagegen können mit dem gebräuchlichen Varianz- resp. regressionsanalytischen F-Test nur ungerichtete Hypothesen über Mittelwerte resp. quadrierte multiple Korrelationskoeffizienten geprüft werden<sup>24)</sup> - zu weiteren Einzelheiten siehe etwa Wainer (1972, 1973) und Gaito (1977a).

über die verschiedenen Arten, statistische Hypothesen zu formulieren und statistisch zu prüfen, ist es im Anschluß an eine Arbeit von Marks (1951) zu einer langandauernden kontroversen Diskussion gekommen, deren seinerzeit vorläufiger Schlußpunkt von Kaiser (1960) gesetzt wurde. Diese und weitere Originalarbeiten haben Lieberman (1971) und Steger (1971) zusammengestellt; einen Überblick über die Diskussion geben auch Glass & Stanley (1970, 288f.); ferner siehe Shaffer (1972) und Gibbons & Pratt (1975).

### *7.5.2 Parametrische und nicht-parametrische Hypothesen und ihre Prüfung*

Formuliert man statistische Hypothesen über die Parameter von Populationsverteilungen, spricht man von „parametrischen Hypothesen“, wobei davon ausgegangen werden muß, daß die Population durch eine endliche Menge dieser Parameter charakterisiert werden kann (vgl. Menges, 1972, 296); Beispiele für derartige Hypothesen finden sich in den Abschnitten 6, 7.2 und 7.3.

Formuliert man dagegen Hypothesen über ganze Verteilungen, ohne diese durch Parameter näher spezifizieren zu können, handelt es sich um „nicht-parametrische Hypothesen“; ein Beispiel wäre etwa:  $H_0: F_1(Y) = F_2(Y)$ , d.h. die abhängige Variable Y hat in den beiden in Frage stehenden Populationen die gleiche Verteilungsfunktion; zu den Einzelheiten siehe etwa Kendall & Stuart (1961, e.g. 161f.) und Menges (1972, 325-328).

Die Prüfung parametrischer Hypothesen erfolgt üblicherweise über „verteilungsgebundene“ Testverfahren, die in der Regel „parametrische“ Tests genannt werden. Das Adjektiv „verteilungsgebunden“ bezieht sich dabei auf die Tatsache, daß die Stichprobenverteilungen der entsprechenden parametrischen Teststatistiken wie  $t$ ,  $\chi^2$  und  $F$  nur unter der Voraussetzung ableitbar sind, daß die Populationsverteilungen der möglichen Rohwerte eine genau spezifizierte Form aufweisen, die allerdings bei der Hypothesenformulierung in der Praxis nur selten expliziert wird. Zur Ableitung der genannten Statistiken muß die Population der Rohwerte durch eine „normal“ genannte Dichte-Funktion beschreibbar sein (zu Einzelheiten siehe etwa Hays, 1977, Kap. 8 bis 11). Sobald diese Annahme nicht aufrecht erhalten werden kann, ist - genau genommen - die Anwendung der parametrischen Tests nicht mehr valide - zur Relativierung dieser Aussage siehe jedoch unten Abschnitt 8.2.

---

<sup>24)</sup> Auf die Bedeutung dieser Aussage kommen wir im Abschnitt 8.2 zurück.



Gegen die Berechtigung dieser Annahme sind verschiedentlich starke Bedenken geäußert worden (u.a. Bradley, 1968, 1972, 1977; Menges, 1972, 248f.). Dies hat zur Entwicklung der sog. „verteilungsfreien“ oder „nicht-parametrischen“ Testverfahren geführt, deren valide Anwendung an weniger strenge oder keine Annahmen über Populationsverteilungen gebunden ist. Das Adjektiv „nicht-parametrisch“ bezieht sich dabei auf die statistischen Hypothesen über Verteilungsfunktionen, die adäquat mittels nicht-parametrischer Tests überprüft werden. Die Bezeichnung „verteilungsfrei“ dagegen soll darauf hindeuten, daß die Ableitung der Stichprobenverteilung der jeweiligen Teststatistik unter Gültigkeit der  $H_0$  unabhängig von spezifischen Annahmen über die Population(en) erfolgt (vgl. zur Unterscheidung neben der unten im Abschnitt 7.5.3.2 genannten Literatur etwa McSweeney & Marascuilo, 1969; sowie McSweeney & Katz, 1978). Allerdings können bestimmte Hypothesen auch mit nicht-parametrischen Tests nur geprüft werden, wenn man bereit ist, Annahmen bzgl. der Populationsverteilung(en) zu treffen, etwa die, daß zwei oder mehr Verteilungen symmetrisch oder sogar von gleicher Form („homomer“) sind - siehe dazu Edgington (1965), Lienert (1973, 107), Marascuilo & McSweeney (1977, 269, 334). Die aus einer WH abgeleitete Hypothese über die Rangordnung von Mittelwerten oder Medianen bspw. kann verteilungsfrei nur geprüft werden, wenn man die Symmetrie der zugrundeliegenden Verteilungen annimmt. Ist man nicht bereit oder in der Lage, derartige Annahmen zu akzeptieren, erlaubt die Ablehnung einer verteilungsfrei geprüften  $H_0$  meist nur die Aussage, daß die untersuchten Stichproben Populationen mit verschiedenen Verteilungsfunktionen entstammen.

Die Ableitung der Stichprobenverteilungen erfolgt bei den meisten nicht-parametrischen Testverfahren aufgrund vglw. einfacher kombinatorischer und wahrscheinlichkeitstheoretischer Überlegungen. Dies wird besonders deutlich, wenn man als Untergruppe der verteilungsfreien Verfahren zur Prüfung nicht-parametrischer Hypothesen die sog. „Randomisierungstests“ betrachtet.

Bei ihnen erhält man bspw. die Prüfverteilung der interessierenden Statistik, indem man alle möglichen Permutationen („Randomisierungen“) oder Zufallsanordnungen der empirisch erhobenen Daten herstellt und für jede dieser Anordnungen den Wert der betr. Statistik errechnet. Auf diese Weise kann eine Häufigkeitsverteilung der Werte der Statistik konstruiert werden, aus der wiederum die relativen Häufigkeiten interessierender extremer Ereignisklassen ermittelt werden. Sind diese bestimmt, verläuft die statistische Signifikanzbeurteilung nach dem im Abschnitt 7.2 skizzierten Verfahren. Bzgl. der Einzelheiten verweisen wir auf Kempthorne (1955), Fisher (1956), Scheffé (1959, 291-330), McHugh (1963), Edgington (1964b, 1969b), Ray (1966), Bradley (1968, Kap. 4) sowie Pfanzagl (1978, 142-147, 150-153); zur Kritik dieses Ansatzes siehe etwa Witte (1980, 122f.).

Ein wesentlicher Nachteil der Permutations- oder Randomisierungstests besteht in dem fast stets enormen Rechenaufwand, den es erfordert, aus allen möglichen Permutationen der Daten die überhaupt möglichen Realisierungen der Teststatistik zu berechnen

(vgl. Edgington, 1964b, 447). Als Ausweg empfiehlt es sich, statt aller möglichen Permutationen der *Rohwerte* lediglich die Permutationen der *Ränge* zu betrachten, die man den Rohwerten zuweisen kann. Auf diesem Prinzip beruhen die gebräuchlichsten verteilungsfreien Verfahren wie die Rangtests nach Wilcoxon, Mann und Whitney, Friedman sowie Kruskal und Wallis (vgl. dazu Lienert, 1973; Marascuilo & McSweeney, 1977).<sup>25)</sup>

Weitere Vereinfachungen im Gebrauch der Randomisierungstests ergeben sich, wenn man die gebräuchlichen parametrischen Verfahren als approximative Randomisierungstests interpretiert (vgl. Edgington, 1966, 1973; Alf & Abrahams, 1972, 1973; Bredenkamp, 1972, 28-33). Diese Interpretation ist dann vertretbar, wenn der Nachweis gelingt, daß die aus allen Permutationen der Daten eines beliebigen Experiments berechenbaren Werte der gewählten Teststatistik approximativ den entsprechenden theoretischen (kontinuierlichen) Stichprobenverteilungen folgen. Dieser Nachweis ist für den t-Test und verschiedene Varianz- und kovarianzanalytische F-Tests in der Tat gelungen (vgl. zusammenfassend Scheffé, 1959, 291-330; ferner Baker & Collier, 1966, 1968; Collier & Baker, 1966; Toothaker, 1971, 1972; sowie Robinson, 1973a, b).

Da für die valide Anwendung von Randomisierungstests nur die Forderung erfüllt sein muß, daß die Untersuchungseinheiten den experimentellen Bedingungen zufällig zugeteilt worden sind, nicht jedoch, daß sie Zufallsstichproben darstellen, ergeben sich aus diesen Befunden wesentliche Konsequenzen bzgl. der Berechtigung parametrischer Tests, auf die wir im Abschnitt 8.2.6 eingehen.

### 7.5.3 Zur Wahl zwischen parametrischen und nicht-parametrischen Verfahren

Aufgrund des bisher Gesagten könnte der Eindruck entstehen, als läge die Entscheidung zwischen den unterschiedlichen Arten der Hypothesenformulierung und -prüfung ausschließlich beim E, zumal aus der WH fast stets parametrische wie nicht-parametrische Hypothesen ableitbar sind. Dieser Eindruck wäre jedoch falsch.

Vielmehr sind die folgenden übergeordneten Entscheidungsgesichtspunkte von zentraler Bedeutung:

- (1) Eine Voraussetzung für die sinnvolle Interpretation der empirischen Resultate besteht darin, daß sowohl die statistischen Hypothesen wie das Auswertungsverfahren dem Skalenniveau der empirischen AV angemessen sind (vgl. Abschnitt 2.4). Wir stellen in den folgenden Abschnitten 7.5.3.1 bis 7.5.3.3 einige Hinweise auf adäquate Auswertetechniken in Abhängigkeit vom Skalenniveau der AV zusammen.

<sup>25)</sup> Ein anderer Ausweg besteht darin, eine Zufallsstichprobe von Werten der Statistik zu ziehen, aufgrund derer man zur Bestimmung der relativen Häufigkeit bzw. Wahrscheinlichkeit der Resultatsklasse von Ereignissen gelangt, die das eine tatsächliche Resultat enthält; man spricht dann von „approximativen Randomisierungstests“ (Edgington, 1964b, 1969a, b, 152-157; Bredenkamp, 1972, 30f.).

- (2) Hat man sich für eine parametrische Hypothese entschieden, ist zu beachten, daß die entsprechenden Tests in der Regel auf (sehr) restriktiven Annahmen beruhen, von denen wir die der normalverteilten Rohwerte in den Populationen bereits angesprochen haben; genauere Einzelheiten zu dieser und den übrigen Voraussetzungen sowie zu den Konsequenzen ihrer Verletzung finden sich im Abschnitt 8.2.
- (3) Unter dem Kriterium der Strenge einer Prüfung ist es wichtig, sich mit der Effizienz der nicht-parametrischen im Vergleich zu den parametrischen Testverfahren zu befassen - dies geschieht im Abschnitt 7.5.4.

Befassen wir uns zunächst mit der Wahl von Tests in Abhängigkeit vom Skalenniveau!

### 7.5.3.1 Auswertung von Häufigkeitsdaten (Nominal-Niveau)

Obwohl davon auszugehen ist, daß in einem *Experiment* relativ selten nominale (qualitative, kategoriale oder Häufigkeits-)Daten erhoben werden, geben wir hier einige Hinweise auf entsprechende Auswertetechniken, weil sie in der nicht-experimentellen Forschung vglw. häufig anzutreffen sind und weil ihr Bekanntheitsgrad bei Psychologen recht gering zu sein scheint.

Die Erhebung von Häufigkeitsdaten (vgl. zur eingehenderen Charakterisierung und Differenzierung etwa Lienert, 1973, 86f.) dient in der Regel der Überprüfung von Hypothesen über die Zusammenhänge zwischen qualitativen Merkmalen. Liegen derartige Zusammenhänge vor, äußern sie sich in unterschiedlichen Häufigkeiten oder Proportionen für einzelne oder mehrere Merkmalskombinationen - man spricht dann von einer „statistischen Interaktion oder Assoziation zwischen qualitativen Merkmalen“, bei deren Vorliegen die einzelnen Merkmale nicht unabhängig voneinander sind.

In der statistischen Null-Hypothese wird allerdings von der statistischen Unabhängigkeit ausgegangen, und unter ihrer Gültigkeit lassen sich erwartete Häufigkeiten oder Proportionen berechnen. Zur Prüfung dieser Hypothese werden häufig die Binomial- oder die Multinomialverteilung sowie deren Spezialfälle herangezogen, und zwar in Abhängigkeit davon, ob ein dichotomes oder polytomes Merkmal vorliegt. Diese Wahrscheinlichkeitsverteilungen können unter bestimmten Voraussetzungen von der  $\chi^2$ -Verteilung approximiert werden.

Die Verfahren zur Auswertung nominaler Daten, die häufig in Form sog. „Kontingenztafeln“ zusammengestellt werden, haben in den letzten Jahren stark an Vielfalt zugenommen; einen zusammenfassenden Überblick über diese Entwicklung, die besonders durch Arbeiten aus dem Bereich der Soziologie begünstigt wurde, findet man bei Meredith, Frederiksen & McLaughlin (1974) sowie besonders bei Smith (1976b). Darüber hinaus sind einzelne Verfahren ausführlicher dargestellt etwa bei Grizzle, Starmer & Koch (1969), Fleiss (1973), Shaffer (1973), Goodman (1978), Upton (1978), Küchler (1979) sowie Langeheine (1980); auf die weitergehenden Literaturhinweise bei den beiden letztgenannten Autoren sei besonders verwiesen.

Als Parallelentwicklung zu dem Log-linearen Ansatz von Goodman (siehe zusammenfassend Goodman, 1978) kann im deutschen Sprachraum die „Konfigurationsfrequenz-

analyse" und ihre Spezialfälle von Krauth & Lienert (1973; vgl. auch Lienert, 1978) verstanden werden, die sich besonders zur Behandlung „klinischer Fragestellungen“ eignet; zum Vergleich der KFA mit dem Log-linearen Ansatz siehe Krauth (1980).

Ferner sei noch der informationstheoretische Ansatz zur Auswertung von Kontingenztabellen erwähnt, der von Kuhback und seinen Mitarbeitern entwickelt worden ist (siehe etwa Ku & Kuliback, 1968; Gokhale & Kuliback, 1978; ferner Adam & Enke, 1972).

Weitere Auswertungshinweise sind den Standard-Lehrbüchern der Statistik sowie insbesondere der im folgenden Abschnitt genannten Literatur zu entnehmen.

### 7.5.3.2 Auswertung von Rangdaten (Ordinal-Niveau)

Ordinale oder Rangdaten unterscheiden sich von nominalen oder Häufigkeitsdaten dadurch, daß sie eine Rangordnung repräsentieren (vgl. zu näheren Einzelheiten u.a. Lienert, 1973, 87-91); dies ist etwa bei Präferenzurteilen der Fall. Andererseits kann man metrische Daten in eine Rangreihe transformieren. Dies ist bspw. dann notwendig, wenn infolge der Verletzung von Annahmen eine parametrische Hypothese durch eine nicht-parametrische ersetzt werden muß, die mittels eines Rangtests voraussetzungsärmer geprüft werden kann. Wenn man bereit ist, bestimmte Annahmen etwa über die Symmetrie der Populationsverteilungen zu akzeptieren, kann eine Hypothese über Mittelwerte bspw. durch eine solche über Mediane ersetzt werden. Die entsprechende Prüfung stellt ein nicht-parametrisches Homologon der Prüfung der Mittelwertshypothese dar, wenn ein Testverfahren gewählt wird, das besonders sensitiv auf Lageunterschiede „anspricht“; es ist dann bspw. der t-Test durch den U-Test von Mann & Whitney (1947; vgl. dazu auch Berchtold, 1979) oder den Wilcoxon-Test (vgl. Lienert, 1973) zu ersetzen, während die Rangvarianzanalysen nach Kruskal & Wallis (1952) und Friedman (1937) einfachen Varianzanalysen entsprechen. Wie insbesondere D'Agostino (1972) und Silverstein (1974) ausführen, sind diese Verfahren einander jeweils „asymptotisch äquivalent“.

Nähere Einzelheiten zu den bei Vorliegen einer Rangskala angemessenen Verfahren entnehme man den Büchern von Siegel (1956, 1976), Walsh (1962, 1965, 1968), Bradley (1968), Gibbons (1971), Puri & Sen (1971), Hollander & Wolfe (1973), Lienert (1973, 1975, 1978), Lehmann (1975), Renn (1975), Marascuilo & McSweeney (1977) sowie Büning & Trenkler (1978), die auch zahlreiche Hinweise zur Auswertung nominaler Daten enthalten; eine zusammenfassende Literatur-Übersicht gibt Singer (1979).

Spezielle Abhandlungen zur Gruppe der „Trendanalysen“ genannten Verfahren finden sich etwa bei Ferguson (1965), Sarris (1968), Marascuilo & McSweeney (1967, 1977) und Bredenkamp (1971).

### 7.5.3.3 Auswertung von Intervalldaten (Intervall-Niveau)

Die Behandlung der Verfahren, die bei Vorliegen intervallskalierter Daten sinnvoll und am teststärksten sind, nimmt in fast jedem Standardlehrbuch den breitesten Raum ein. Wir können uns daher an dieser Stelle darauf beschränken, nur einige derjenigen (Lehr-)Bücher anzugeben, die wertvolle Hinweise zur Auswertung von Experimenten geben können. Bei der Auswahl haben wir uns auch an dem Kriterium der raschen Verfügbarkeit dieser Werke orientiert.

Die umfassendste Darstellung der verschiedensten Aspekte der Versuchsauswertung findet sich derzeit wohl in der „Verfahrensbibliothek Versuchsplanning und -auswertung“ von Rasch, Herrendörfer, Bock & Busch (1978); vgl. ergänzend dazu Bliss (1967, 1970) und Bortz (1979). Weniger umfassend im Detail, aber stärker die Verbindung zwischen Auswerteverfahren und wissenschaftlicher Hypothese betonend sind u.a. Namboodiri, Carter & Blalock (1975), Box, Hunter & Hunter (1978) sowie Henning & Muthig (1979). überwiegend mit der Auswertung der zahlreichen „klassischen“ Versuchspläne der Varianzanalyse befassen sich neben Winer (1962, 1971), Edwards (1971, 1980) und Keppel (1973) u.a. auch Lindquist (1953), Cox (1958), Cochran & Cox (1968), Mendenhall (1968), Kirk (1968), Myers (1972), Snedecor & Cochran (1972), Kempthorne (1973), Lindman (1974), Lee (1975), John & Quenouille (1977), Eimer (1978), Diehl (1979) und McGuigan (1979).

Auf die Behandlung der gleichen einfachen und komplexen Designs durch regressionsanalytische Verfahren gehen u. a. Draper & Smith (1966), Kerlinger & Pedhazur (1973), Cohen & Cohen (1975), Gaensslen & Schubö (1973, 1976) und Moosbrugger (1978) ein.

Die parametrischen Trendanalysen werden darüber hinaus gesondert von Gaito & Turner (1963), Bredenkamp (1968), Hubert (1973), Gaito (1977b) und Cohen (1980) behandelt.

Über multivariate Testverfahren informieren Cramer & Bock (1966), Morrison (1967, 1976), Bock & Haggard (1968), Rulon & Brooks (1968), Tatsuoka (1969, 1971), McCall (1970), Cooley & Lohnes (1971), Overall & Klett (1972), Gaensslen & Schubö (1973, 1976), Kerlinger & Pedhazur (1973), Finn (1974), Bock (1975), Cohen & Cohen (1975), Harris (1975), Woodward & Overall (1975), Moosbrugger (1978) sowie Bortz (1979).

Unabhängig von der individuellen Entscheidung für oder gegen eine der vorstehenden Arbeiten über parametrische Hypothesentestung halten wir eine zusätzliche Beschäftigung mit dem Konzept der „Exploratorischen Daten-Analyse“, wie es von Tukey (1977; vgl. auch Mosteller & Tukey, 1977) entwickelt wurde, für gewinnbringend.

#### 7.5.4 Zur Frage der relativen Effizienz

Wenn man der Frage nachgeht, warum in der Praxis fast grundsätzlich parametrische Tests den nicht-parametrischen vorgezogen werden, stößt man auf zahlreiche Aspekte, unter denen man die beiden Gruppen von Verfahren vergleichen kann - entsprechende Erörterungen findet man etwa bei Siegel (1956, 1976), Gaito (1959b), Bradley (1968, Kap. 2, 1972) und bei Lienert (1973, Kap. 4); zur Kritik der Unterscheidung siehe etwa McSweeney & Marascuilo (1969).

Wir wollen uns hier auf einige wenige Punkte beschränken. Zum einen liegt die Bevorzugung daran, daß die parametrischen Verfahren vielseitiger einsetzbar sind und in einigen Fällen sogar die einzige Möglichkeit zur Auswertung darstellen; letzteres gilt insbesondere bei der Prüfung komplexer Hypothesen, etwa über (varianzanalytische) statistische Interaktionen, oder auch bei der Auswertung komplexer, insbesondere multivariater Versuchspläne - siehe hierzu im einzelnen Bradley (1968), Puri & Sen (1971, e.g. 331-337), McSweeney & Katz (1978) und Singer (1979).

Zum anderen stellen die gebräuchlichen parametrischen Tests die teststärksten und insgesamt „besten“ Tests „ihrer“ Hypothesen dar, wenn alle Voraussetzungen zu ihrer Anwendung erfüllt sind (vgl. zu den Test-Gütekriterien Menges, 1972, 333-335). Unter genau dieser Bedingung haben die parametrischen Tests eine Effizienz von Eins, während die analogen nicht-parametrischen Tests fast ausnahmslos<sup>26)</sup> weniger effizient, d.h. - vereinfacht ausgedrückt - weniger teststark sind (siehe etwa Büning & Trenkler, 1978, 282). Exakte Definitionen der verschiedenen Arten von „(relativer) Effizienz“ findet man bei Marascuilo & McSweeney (1977, Kap. 4) und bei Büning & Trenkler (1978, Kap. 9).

Praktisch bedeutet eine geringere Effizienz, daß unter sonst gleichen Bedingungen bei Verwendung eines nicht-parametrischen Tests mehr Vpn benötigt werden als beim analogen parametrischen Test, um eine zutreffende Alternativhypothese aufgrund der Stichprobendaten annehmen zu können.

Sind dagegen mehrere Voraussetzungen zur validen Anwendung der parametrischen Verfahren simultan verletzt, liegt deren Effizienz in der Regel beträchtlich *unter* der der homologen nicht-parametrischen Tests - vgl. zu Einzelheiten Hodges & Lehmann (1956), Wetherill (1960), Pratt (1964), Renn (1975, 43f.), Blair, Higgins & Smitley (1980) und Hager et al. (im Druck).

---

<sup>26)</sup> Eine der wesentlichen Ausnahmen stellen die sog. „Normal-Scores-Tests“ dar, die stets mindestens so effizient sind wie die entsprechenden parametrischen Tests (vgl. Bradley, 1968, Kap. 6; Lienert, 1973, 257-262; Marascuilo & McSweeney, 1977, Kap. 11 und 12).

Daneben ist für eine ganze Reihe der häufiger verwendeten nicht-parametrischen Testverfahren gezeigt worden, daß ihre „relative asymptotische Effizienz“ (siehe Pitman, 1948; Bradley, 1968, 56-62; Büning & Trenkler, 1978, 275-282) selbst dann, wenn alle Voraussetzungen für parametrische Tests erfüllt sind, nicht unter einen - häufig nahe bei Eins liegenden - Minimalwert absinkt (vgl. u.a. Hodges & Lehmann, 1956, und Berchtold, 1979).

Die Werte für die relative asymptotische Effizienz findet man insbesondere in den Büchern von Bradley (1968, 60f.), Lienert (1973, 1978), Marascuilo & McSweeney (1977, 87) sowie Büning & Trenkler (1978, 282). Diese Hinweise sind insofern wichtig, als die Teststärkenbestimmung sich bei den meisten Rangverfahren überaus schwierig gestaltet, so daß man in diesen Fällen darauf angewiesen ist, diese unter Verwendung der nicht-zentralen Verteilungen der homologen parametrischen Tests und der Maßzahl der relativen asymptotischen Effizienz ungefähr abzuschätzen. Wie wir bereits mehrfach erwähnt haben, ist die Kenntnis und Kontrolle der Teststärke eine notwendige Voraussetzung für eine strenge Prüfung einer WH.

## 7.6 Zusammenfassung

Im Teil 6 haben wir herausgearbeitet, daß die Prüfung einer wissenschaftlichen Hypothese i. a. nur über eine aus ihr abgeleitete statistische Hypothese erfolgen kann. Eine der Möglichkeiten, eine statistische Hypothese zu beurteilen, bietet der Signifikanztest. Der vorangegangene Teil 7 diente im wesentlichen der Darstellung einiger Probleme, die mit der üblichen Anwendung von Signifikanztests verbunden sind und die dazu führen, daß die Beurteilung der WH häufig nicht angemessen möglich ist.

Demzufolge wurden einige Modifikationen an der „Anwenderversion“ des Signifikanztests aufgegriffen, die auf die Kontrolle seiner Determinanten hinauslaufen, und es wurde eine vorläufige Zielvorstellung bzgl. der Anwendung von Signifikanztests vorgestellt, die in den Teilen 10 und 11 vertieft werden soll.

Ferner wurde aufgezeigt, daß unterschiedlichen Arten von statistischen Hypothesen auch unterschiedliche Prüfverfahren entsprechen, deren valide Anwendung teils an sehr restriktive und teils an nur schwache Voraussetzungen gebunden ist. Auf die strengen Voraussetzungen zur Anwendung parametrischer Tests gehen wir im Abschnitt 8.2 ein, dem einige Ausführungen über die Beziehungen zwischen der WH und der statistischen Hypothese vorausgehen werden.

## 8. Störfaktoren der statistischen Validität und ihre Ausschaltung

Um fälschliche Falsifikationen und ungerechtfertigte Bewährungen der wissenschaftlichen Hypothese zu vermeiden, müssen bei der Prüfung der aus ihr abgeleiteten statistischen Hypothesen die Fehlerwahrscheinlichkeiten erster und zweiter Art möglichst gering sein (Teil 6). Alle Faktoren, die letztlich dazu führen, daß diese Fehlerwahrscheinlichkeiten  $\alpha$  und  $\beta$  erhöht werden, setzen die statistische Validität der Untersuchung herab. Wir wollen die wichtigsten dieser Störfaktoren der statistischen Validität im folgenden -wieder in Gruppen zusammengefaßt - besprechen und jeweils auch Maßnahmen zu ihrer Vermeidung erörtern. Diese Störfaktoren (StatV) umfassen die u.E. schwerstwiegenden Fehler, die bei der statistischen Auswertung empirischer Untersuchungen (insbesondere Experimente) gemacht werden können und leider auch immer wieder gemacht werden.

### 8.1 Falsche statistische Hypothesen und Verfahren

#### 8.1.1 Die wichtigsten Beziehungen zwischen psychologischen und statistischen Hypothesen

Eine wissenschaftliche (Kausal-)Hypothese WH kann über die Prüfung einer statistischen Hypothese SH falsifiziert werden, wenn zwischen ihnen eine logische Implikationsbeziehung besteht, d.h. wenn gilt:  $WH \rightarrow SH$  (vgl. Teil 6). Wir wollen jetzt verschiedene Arten von psychologischen Kausalhypothesen betrachten und dabei jeweils untersuchen, welche statistischen Hypothesen aus ihnen ableitbar sind und mit welchem statistischen Verfahren diese geprüft werden können. Aus diesen Darstellungen werden sich einige mögliche Fehler ergeben, die die statistische Validität herabsetzen (zu weiteren Einzelheiten siehe Hager & Westermann, im Druck, a).

(1) Aus unserer Beispielhypothese ließ sich ableiten, daß bei Vorliegen von Dissonanz die Werte auf der Einstellungsvariablen höher sind als ohne Dissonanz. Aus  $WH_u$  folgte deshalb eine statistische Hypothese wie  $H_1: \mu_2 - \mu_1 > 0$ . In dieser Beziehung ist  $WH_u$  typisch für Kausalhypothesen in der Psychologie: In der Regel implizieren psychologische Kausalhypothesen *gerichtete* statistische Alternativhypothesen (Bredenkamp, 1972; Hager & Westermann, im Druck, a). Dies hat wichtige Konsequenzen für die Durchführung von Signifikanztests (a. a.o.): Damit die übergeordnete wissenschaftliche Kausalhypothese überhaupt falsifizierbar ist, muß man sich für die Falschheit der statistischen Alternativhypothese (und damit für die Richtigkeit der  $H_0$ ) entscheiden können. Diese Entscheidung ist aber für den Experimentator nur dann zu verantworten, wenn er die Wahrscheinlichkeit dafür kennt, daß diese Entscheidung falsch ist und wenn diese Fehlerwahrscheinlichkeit ihm klein genug



erscheint. Zur Prüfung von wissenschaftlichen Hypothesen, die eine statistische Alternativhypothese implizieren, muß also auch  $\beta$  auf einen kleinen Wert festgelegt werden.

(2) Läßt sich aus einer wissenschaftlichen Hypothese nur ableiten, daß Mittelwertsunterschiede in irgendeiner Richtung bestehen, folgt aus ihr eine ungerichtete Alternativhypothese wie  $H_1 : \mu_2 - \mu_1 \neq 0$ . Wissenschaftlich dürften solche unspezifischen Vorhersagen nur von geringem Interesse sein.

(3) Relativ selten sind auch Kausalhypothesen, deren ausschließliche Aussage darin besteht, daß unter bestimmten Bedingungen keine Unterschiede oder Veränderungen zu erwarten sind. Aus ihnen folgen einfache statistische Nullhypothesen wie  $H_0 : \mu_1 - \mu_2 = 0$ .

(4) In den exakten Naturwissenschaften sind Theorien im allgemeinen so präzise formuliert, daß im Konklusionsteil („dann . . .“) von Kausalhypothesen genaue Angaben über den Wert einer bestimmten Variablen gemacht werden. Aus ihnen folgen dann statistische Nullhypothesen, nach denen ein Parameter einen ganz bestimmten Wert hat, also z.B.  $H_0 : \mu = 45$ . In der Psychologie sind solch präzise Folgerungen nur in den seltenen Fällen zu erwarten, in denen Theorien nicht nur verbal, sondern auch mathematisch formuliert sind (vgl. Restle & Greeno, 1970; Bredenkamp, 1972, 186-198; Coombs, Dawes & Tversky, 1975; Bredenkamp & Hager, 1979).

In den bisher besprochenen vier Fällen impliziert die wissenschaftliche Hypothese stets eine statistische Hypothese über einen oder zwei Mittelwerte. Kann man bestimmte Annahmen über die Verteilung der Werte in den Populationen machen, können die jeweiligen statistischen Null-Hypothesen über den parametrischen t-Test geprüft werden. Kann oder will man die Verteilungsannahmen nicht aufrechterhalten, sind aus der psychologischen Hypothese andere statistische Hypothesen (z.B. über Mediane) abzuleiten, die über nicht-parametrische Testverfahren geprüft werden können.

Als wichtiges Ergebnis unserer Überlegungen ist festzuhalten, daß in den meisten Fällen der psychologischen Forschungspraxis aus der wissenschaftlichen Hypothese eine gerichtete Alternativhypothese folgt, deren Prüfung über einen einseitigen statistischen Test erfolgen sollte. Zweiseitige Tests sind nur dann gerechtfertigt, wenn von der psychologischen Hypothese - ausnahmsweise - tatsächlich eine ungerichtete Alternativhypothese oder aber eine einfache Null-Hypothese impliziert wird - wir kommen darauf im Teil 11 zurück.

Wenden wir uns nun weiteren Arten von psychologischen Hypothesen zu!

(5) Schon häufiger als die in (4) besprochenen exakten numerischen Vorhersagen treten in der Psychologie Hypothesen auf, die ganz bestimmte *funktionale*

Zusammenhänge zwischen Variablen annehmen (Beispiel: die „psychophysischen Funktionen“). Die aus ihnen ableitbaren statistischen Nullhypothesen werden durch parametrische oder nicht-parametrische Trendanalysen geprüft.

(6) Wissenschaftliche Kausalhypothesen können so formuliert sein, daß sie die Gleichheit mehrerer Parameter implizieren; allerdings ist diese Art von Kausalhypothesen in der Psychologie noch seltener anzutreffen als der unter (3) dargestellte Sonderfall, aus dem die Gleichheit zweier Parameter folgte. In der Mehrzahl der Fälle handelt es sich bei den in Frage stehenden Parametern um Mittelwerte. Deshalb kann die statistische Null-Hypothese wie folgt angegeben werden :

$$H_0: \mu_1 = \mu_2 = \dots = \mu_K.$$

Es handelt sich hierbei um die Null-Hypothese der parametrischen Varianzanalyse. Die zu ihr gehörige Alternativhypothese  $H_1$  besagt, daß zwischen mindestens zwei Mittelwerten ein Unterschied besteht; sie lautet also formal:

$$H_1: \mu_m \neq \mu_{m'}, \text{ für mindestens ein Paar von Bedingungen } X_m \text{ und } X_{m'}, \text{ wobei gilt, daß } m \neq m'.$$

Diese Art von Alternativhypothesen folgt aus psychologischen Kausalhypothesen, die unspezifisch irgendeinen Unterschied in irgendeiner Richtung vorhersagen. Gerade wegen des ausgeprägten Mangels an Spezifität der Vorhersage sind solche wissenschaftlichen Hypothesen nur von geringem Interesse.

Insgesamt gesehen folgen also aus einer wissenschaftlichen Hypothese der Psychologie nur selten die Null- oder Alternativhypothese einer *Varianzanalyse*. Das steht im Gegensatz zur Tatsache, daß bei der großen Mehrheit der veröffentlichten experimentellen Untersuchungen die Auswertung über parametrische und nicht-parametrische Varianzanalysen erfolgt und daß die Bücher zur Versuchsplanung („Experimental Design“) überwiegend varianzanalytische Auswertungstechniken behandeln - siehe dazu die Literaturangaben in den Abschnitten 7.5.3.2 und 7.5.3.3.

Man kann also davon ausgehen, daß das angewendete statistische Verfahren in vielen Fällen gar nicht die Information erbringt, die zur (strengen) Prüfung der betrachteten Kausalhypothese erforderlich ist (Hager & Westermann, im Druck, a).

(7) Psychologische Hypothesen und Theorien sind nur in den seltensten Fällen präzise genug, um die Ableitung numerischer Vorhersagen oder Aussagen über die Form funktionaler Beziehungen vornehmen zu können. Typischer für den Entwicklungsstand der meisten psychologischen Teilgebiete sind wissenschaftliche Hypothesen wie folgende Ergänzung  $WH_v$  zu unserer Kausalhypothese  $WH_u$ : „Je stärker die auftretende Dissonanz ist, desto größer ist die

Einstellungsänderung in Richtung auf die dissonanzerzeugende Kognition.“ Aus derartigen wissenschaftlichen Hypothesen folgt stets eine bestimmte *Rangordnung von Mittelwerten*, also z. B.:

$$H_1: \mu_2 < \mu_3 < \dots < \mu_K$$

mit der dazugehörigen Null-Hypothese

$$H_0: \mu_m \geq \mu_{m'}, \text{ für mindestens ein Paar } m \text{ und } m', \text{ mit } m < m'.$$

Die Planung von Experimenten zur Prüfung solcher Hypothesen über einen monotonen Trend wirft einige schwierige Probleme auf, die wir im Abschnitt 8.3.2 ansprechen werden.

(8) Aus psychologischen Theorien können häufig Hypothesen abgeleitet werden, die einander ergänzen wie die oben besprochenen  $WH_u$  und  $WH_v$ . In diesem Fall können beide simultan in einem einzigen Experiment überprüft werden. Dazu müssen - um im Beispiel zu bleiben - eine experimentelle Bedingung „Keine Dissonanz“ ( $X_1$ ) und mindestens zwei Treatments mit ansteigender Stärke der Dissonanz ( $X_2, X_3, \dots, X_K$ ) hergestellt werden. Die Hypothese  $WH_u$  impliziert dann eine statistische Hypothese wie die folgende  $H_{1(1)}$ :

$$H_{1(1)}: \frac{1}{K-1} \left\{ (\mu_2 + \mu_3 + \dots + \mu_K) \right\} > \mu_1;$$

die Hypothese  $WH_v$  dagegen impliziert die  $H_{1(2)}$ :

$$H_{1(2)}: \mu_2 < \mu_3 < \dots < \mu_K.$$

Auf die Prüfung solcher (u.U. hierarchisch geordneter) Mengen von Hypothesen gehen wir im Abschnitt 8.3.1 ein.

(9) Lassen sich aus einer psychologischen Theorie mehrere statistische Hypothesen ableiten, die sich auf die Werte der gleichen AV unter den Ausprägungen verschiedener UV beziehen, können diese Hypothesen über mehrfaktorielle Designs simultan überprüft werden. Diese Designs gestatten auch die Prüfung von Hypothesen über die Wechselwirkung (Interaktion) der unabhängigen Variablen in bezug auf die abhängige Variable. Eine statistische Interaktionshypothese würde z.B. aus einer wissenschaftlichen Hypothese abzuleiten sein, die folgendes aussagt: „Kognitive Dissonanz führt zu einer Einstellungsänderung in Richtung auf die Information aus einer glaubwürdigen Quelle, aber in entgegengesetzter Richtung bei Information aus einer unglaublichen Quelle.“ Derartigen Interaktionshypothesen wird ein zunehmendes Interesse entgegengebracht. Auf die mit ihrer Prüfung verbundenen Probleme gehen wir im Abschnitt 8.5 ein.

Wir haben in den neun Punkten die wichtigsten Grundformen psychologischer Kausalhypothesen dargestellt, und wir haben erörtert, welche Arten von statistischen Hypothesen jeweils aus ihnen folgen. Ferner sind wir darauf eingegangen, mit welchen Verfahren die statistischen Hypothesen geprüft werden können. Aus dieser Diskussion ergeben sich unmittelbar drei mögliche Beeinträchtigungen der statistischen Validität (StatV) einer Untersuchung, denen die Abschnitte 8.1.2 bis 8.1.4 gewidmet sind.

### *8.1.2 Falsche Umsetzung der wissenschaftlichen in eine statistische Hypothese als Störfaktor (StatV)*

Die Prüfung einer statistischen Hypothese SH kann nur dann der strengen Prüfung einer wissenschaftlichen Kausalhypothese WH dienen, wenn SH tatsächlich ein Implikat von WH ist. Dieser Tatbestand wird in der Forschungspraxis sehr häufig nicht beachtet; dies äußert sich bspw. darin, daß eine Varianzanalyse durchgeführt wird, obwohl aus der psychologischen Hypothese statistische Hypothesen über die Ordnung zweier oder mehrerer Mittelwerte ableitbar sind (vgl. die Punkte (1), (6) und (7) im Abschnitt 8.1.1). Diese Vorgehensweise setzt die Strenge einer Prüfung herab, wie wir am Beispiel aufzeigen wollen (vgl. auch Hager & Westermann, im Druck, a): Aus den im Punkt (7) besprochenen psychologischen Hypothesen  $WH_u$  und  $WH_v$  folgt die statistische Hypothese:

$$H_1: \mu_1 < \mu_2 < \mu_3, \text{ mit } H_0: \mu_1 \geq \mu_2 \text{ und/oder } \mu_2 \geq \mu_3.$$

Zum Vergleich: Die entsprechenden Hypothesen einer einfachen Varianzanalyse lauten

$$H'_0: \mu_1 = \mu_2 = \mu_3 \text{ und } H'_1: \mu_1 \neq \mu_2 \text{ und/oder } \mu_2 \neq \mu_3.$$

Die Alternativhypothese der Varianzanalyse umfaßt also auch „wahre“ Parameterverhältnisse, die im Widerspruch zur psychologischen Hypothese stehen (beispielsweise  $\mu_1 > \mu_2 = \mu_3$ ). Wird also fälschlicherweise aus den zu prüfenden wissenschaftlichen Hypothesen  $WH_u$  und  $WH_v$  statt der obigen  $H_1$  die Alternativhypothese einer Varianzanalyse abgeleitet, so ist (unter sonst gleichen Bedingungen) für den Fall der Falschheit von  $WH_u$  und  $WH_v$  die Wahrscheinlichkeit hypothesenkonträrer Ergebnisse geringer. Durch inadäquate Ableitung statistischer aus wissenschaftlichen Hypothesen wird also die Strenge des Prüfexperiments herabgesetzt (a. a. 0.).

### 8.1.3 Falsche Auswahl der zu prüfenden statistischen Hypothese

In der Regel können aus einer psychologischen Hypothese mehrere statistische Hypothesen abgeleitet werden. Die Prüfung der wissenschaftlichen Hypothese ist dann am strengsten, wenn sie über diejenige der aus ihr ableitbaren statistischen Hypothesen erfolgt, bei der die Wahrscheinlichkeiten für falsche Bewährungen und falsche Falsifikationen am geringsten ist (siehe im einzelnen Hager & Westermann, im Druck, a).

Verdeutlichen wir dies an zwei Beispielen!

- (1) Folgt aus der HW eine Rangordnung von Mittelwerten, ist unter sonst gleichen Bedingungen eine empirische Prüfung um so strenger, je mehr Mittelwerte experimentell verglichen werden.
- (2) Die Prüfung psychophysischer Trendhypothesen an den Daten jeder einzelnen Person ist strenger als eine Prüfung an Mittelwerten von verschiedenen Personen - vgl. zu Einzelheiten Bredenkamp (1980, 12, 49f.).

### 8.1.4 Falsche statistische Analyse

Hat man aus der wissenschaftlichen Hypothese eine adäquate statistische Hypothese abgeleitet, kann die Validität einer Untersuchung noch durch die falsche Wahl der statistischen Prüfverfahren herabgesetzt werden. Dieser Fall tritt bspw. dann ein, wenn sich nach dem Experiment herausstellt, daß bestimmte Voraussetzungen bzgl. der validen Anwendung des betr. Tests aufgrund der Daten nicht aufrechterhalten werden können, und wenn dann dieser Test dennoch zur Anwendung gelangt - vgl. hierzu im einzelnen Abschnitt 8.2. Ferner werden sehr häufig statistische Tests benutzt, mit denen nur eine ungerichtete Alternativhypothese beurteilt werden kann, obwohl aus der WH eindeutig eine gerichtete Alternativhypothese folgt (siehe Abschnitt 8.1.1, Punkt (6)). Diese *oft allerdinge unvermeidliche* Verfahrensweise hat zur Folge, daß der Ablehnungsbereich in der einzig interessierenden Richtung um die Hälfte zu gering gewählt wird, wodurch zwangsläufig die Wahrscheinlichkeit für einen Fehler 2. Art mehr als verdoppelt wird, wenn wir davon ausgehen, daß tatsächlich die aus der WH abgeleitete statistische Alternativhypothese zutrifft. Durch die somit insgesamt erhöhten Fehlerwahrscheinlichkeiten bei der Entscheidung über die statistische Hypothese wird auch die Wahrscheinlichkeit für eine falsche Entscheidung über die wissenschaftliche Hypothese erhöht.

Weitere in diesem Zusammenhang mögliche Fehlerquellen ergeben sich aus den Ausführungen in den Abschnitten 7.5, 8.2, 8.4.1 und 8.4.6; siehe ferner Lindquist (1953, 7f.) sowie Mosteller (1968, 122).

## 8.2 Verletzung der Annahmen bei statistischen Tests als Störfaktor (StatV)

Die Fehlerwahrscheinlichkeiten bei der statistischen Hypothesenprüfung und damit auch die Wahrscheinlichkeiten für falsche Entscheidungen über das Zutreffen oder Nicht-Zutreffen der wissenschaftlichen Hypothesen können auch dadurch ansteigen, daß die für die Anwendung der inferenzstatistischen Verfahren notwendigen Voraussetzungen oder Annahmen nicht erfüllt sind.

Während die valide Anwendung von nicht-parametrischen Auswertetechniken häufig nur an vglw. schwache Voraussetzungen gebunden ist, beruhen die parametrischen Tests stets auf mehreren restriktiven Annahmen resp. Voraussetzungen. Wegen der bereits wiederholt angesprochenen zentralen Bedeutung, die den parametrischen Verfahren in der Praxis zukommt, wollen wir diese Voraussetzungen im folgenden etwas ausführlicher darstellen und die Konsequenzen ihrer Verletzung erörtern.

### 8.2.1 Das Allgemeine Lineare Modell (ALM) und die Annahmen

Die meisten parametrischen Hypothesen, die der empirisch arbeitende Wissenschaftler einer Prüfung zu unterziehen beabsichtigt, beziehen sich auf Mittelwerte, Varianzen und Korrelations- bzw. Regressionskoeffizienten. Die zur Prüfung benutzten Test-Statistiken sind in der Regel entweder t-,  $\chi^2$ - oder F-verteilt.

Diese Testverfahren lassen sich allesamt aus einem einzigen grundlegenden Modell ableiten, dem sog. „Allgemeinen Linearen Modell“ (ALM). Dieses ist durch die folgenden Eigenschaften zu kennzeichnen (vgl. dazu im einzelnen Moosbrugger, 1978, 57; ferner Fennessey, 1968, 3; Auslitz, Hesse & Rieder, 1975, 3; Schach & Schäfer, 1978, 5):

- (1)  $Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \dots + \beta_m X_{mi} + \dots + \beta_K X_{Ki} + e_i$
- (2) Der Mittelwert der Fehler ist in jeder der Populationen gleich Null.
- (3) Die Fehler  $e_i$  sind innerhalb der experimentellen Bedingungen wie auch zwischen ihnen unabhängig voneinander.
- (4) Die Varianz  $\sigma_e^2$  der Fehler  $e_i$  ist in jeder der K Populationen gleich oder homogen:  

$$\sigma_{e_1}^2 = \sigma_{e_2}^2 = \dots = \sigma_{e_K}^2 = \sigma_e^2$$

Die Merkmalsausprägung jeder Vp i auf der abhängigen oder Kriteriumsvariablen Y, also  $Y_i$ , wird im ALM also dargestellt als Summe der mit  $\beta_k$  gewichteten Werte  $X_k$  der unabhängigen oder Prädiktorvariablen („Einflußgrößen“) X und einer Zufallsgröße  $e_i$ , dem unsystematischen Fehler oder Residuum. Die Werte  $X_k$  sind (vom E) festgelegte Größen, von denen  $X_0$  in der Regel gleich

Eins ist; die „Gewichtsgrößen“  $\beta_k$  bezeichnen der Größe nach unbekannte Parameter, die aus den Daten nach dem Kriterium der „Kleinsten Quadrate“ zu bestimmen sind (vgl. dazu den Satz von Gauss-Markoff, etwa in Scheffé, 1959, 13-19; oder in Menges, 1972, 319).

Zur Prüfung von Hypothesen vor dem Hintergrund des ALM können zwei im wesentlichen formal unterschiedliche Wege beschritten werden.

Wertet man die empirischen Daten aus  $K \geq 2$  Stichproben über einen Simultant-vergleich der  $K$  Mittelwerte aus, spricht man traditionell von einer „Varianzanalyse“ (VA). (Vgl. Fisher, 1925, 1950; ferner Scheffé, 1959; Hays, 1963, 1977; Cochran & Bliss, 1970, sowie die in Abschnitt 7.5.3 angegebene Literatur.)

Bei der Prüfung der Mittelwerts-Hypothese wird davon ausgegangen, daß die  $K$  Modalitäten des Faktors vom E bewußt ausgewählt und festgelegt worden sind, daß m.a. W. sog. „fixierte Effekte“ vorliegen (vgl. Abschnitt 2.3). Verwendet der E dagegen in seinem Experiment eine Zufallsstichprobe der Größe  $K$  aus einer Population möglicher Abstufungen der UV, so liegen „zufällige Effekte“ vor; mit dem Experiment werden statistische Hypothesen geprüft, die sich auf die Varianz dieser Effekte beziehen. Werden zufällige und fixierte Effekte in einem Experiment simultan untersucht, spricht man von „gemischten Effekten“.

Diese Unterscheidungen sind wichtig, weil ihnen unterschiedliche varianzanalytische Modelle und F-Tests entsprechen - vgl. zu Einzelheiten insbesondere Eisenhart (1947), Wilk & Kempthorne (1955) oder Hays (1977, 377f.). Wir beziehen uns im folgenden stets auf das varianzanalytische Modell der fixierten Effekte - zur Begründung siehe die Abschnitte 2.3 und 8.2.3.

Prüft man dagegen Hypothesen über (multiple) Korrelationsquadrate  $R_{Y.X_k}^2$  (resp. über Regressionskoeffizienten  $\beta_k$ ), nennt man diese Art der Auswertung „Multiple Regressions- und Korrelationsanalyse“ (MRA); zur formalen Unterscheidung siehe im einzelnen etwa Auslitz, Hesse & Rieder (1975, 7f.) sowie Schach & Schäfer (1978, 6).

Die oben erwähnte VA mit fixierten Effekten ist als Spezialfall der allgemeineren (und im ganzen vielseitigeren) MRA darstellbar; Einzelheiten hierzu entnehme man etwa Jennings (1967), Cohen (1968), Darlington (1968), Fennessey (1968), Wottawa (1974), Auslitz, Hesse & Rieder (1975), Schach & Schäfer (1978) sowie der bereits im Abschnitt 7.5.3.3 genannten Literatur; zu bestimmten konzeptuellen Unterschieden siehe etwa Witte (1978, 1980, 172f.).

Diese Auswerteverfahren lassen sich ihrerseits unter dem allgemeinen Modell der „Kanonischen Analyse“ subsumieren, wie Knapp (1978) gezeigt hat.

Zur validen Anwendung der o. gen. Tests und Prüfverteilungen sind neben den bereits unter (1) bis (4) aufgeführten Annahmen noch die folgenden Annahmen zu treffen (vgl. dazu auch Eisenhart, 1947; Cochran, 1947; Lindquist,

1953, 72-78; Gaito, 1959b; Scheffé, 1959, Kap. 10; Moosbrugger, 1978, 69; sowie einführende Lehrbücher der Versuchsplanung und -auswertung):

- (5) Die Terme  $e_i$  sind in jeder der im Experiment untersuchten Populationen normalverteilt:
- (6) Die zu untersuchenden Rohwerte stellen Zufallsstichproben aus den interessierenden Grundgesamtheiten dar.

Bei der Darstellung der sechs Voraussetzungen ist ein univariater einfaktorieller Versuchsplan mit jeweils mehreren Vpn pro experimenteller Bedingung zugrunde gelegt worden; die Voraussetzungen beziehen sich jedoch in modifizierter und/oder erweiterter Form auch auf alle anderen Varianz- und regressionsanalytischen Designs. Auf Erweiterungen dieser Voraussetzungen gehen wir im Zusammenhang mit der Kovarianzanalyse (Abschnitt 8.4.2) und mit wiederholten Messungen (Abschnitt 8.4.6) ein, weil diesen Verfahren in der Praxis eine große Bedeutung zukommt; zu weiteren Einzelheiten verweisen wir auf einschlägige Lehrbücher - siehe dazu etwa Abschnitt 7.5.3.3.

Diese Voraussetzungen sind zur validen Hypothesenprüfung unerlässlich. Zu ihnen treten noch zwei Restriktionen, die mit der Hypothesenprüfung jedoch nicht im Zusammenhang stehen:

- (7) Die sog. „Reparametrisierungsbedingung“ ist erforderlich, um zu mathematisch eindeutigen Lösungen der sog. „Normalgleichungen“ gelangen zu können. Sie besagt für das varianzanalytische Modell der fixierten Effekte (und damit auch für die MRA), daß - in varianzanalytischer Terminologie - die Summe der (einfachen) Abweichungen der Treatment-Mittelwerte  $\mu_k$  von ihrem Gesamtmittelwert  $\mu$  für jeden Faktor gleich Null sein muß.

Für den Praktiker ist diese Restriktion meist ohne Belang, weswegen wir sie nicht erörtern; Einzelheiten sind etwa Mendenhall (1968, Kap. 6), Searle (1971 a, 209-220), Glass, Peckham & Sanders (1972, 241) und Rasch et al. (1978, 65-67) zu entnehmen.

- (8) Die empirische AV soll mindestens intervallskaliert sein, damit die Ergebnisse der parametrischen Auswertung über eine VA oder MRA sinnvoll interpretierbar sind - siehe dazu Abschnitt 2.4.

Im Zusammenhang mit den sechs Voraussetzungen wollen wir im folgenden jeweils drei Fragen nachgehen:

Wie können diese Voraussetzungen geprüft werden?

Ist eine derartige Prüfung notwendig und/oder sinnvoll?

Welche Konsequenzen ergeben sich bei Verletzung der jeweiligen Voraussetzung?



### 8.2.2 Additivität

Im Rahmen der auf dem ALM basierenden Auswertetechniken und insbesondere der „klassischen“ VA wird der Terminus „Additivität“ in zwei Bedeutungen benutzt (vgl. u.a. Gaito, 1959b; Lee, 1961; Glass, Peckham & Sanders, 1972; Henning, 1978).

- (1) Zum einen bezieht er sich auf die additive Verknüpfung der Komponenten, die jeden einzelnen Rohwert konstituieren; etwa:

$$(8.1) \quad \begin{aligned} Y_{ik} &= \mu_k + \beta_k + e_{ik} \text{ (einfaktorielle VA)} \\ &= \beta_0 + \beta_1 X_1 + e_{ik} \text{ (regressionsanalytische Darstellung);} \end{aligned}$$

oder für den Fall einer univariaten bifaktoriellen Versuchsanlage mit zwei UV A und B sowie einer AV Y:

$$(8.2) \quad \begin{aligned} Y_{ijk} &= \mu + \beta_k + \alpha_j + (\alpha\beta)_{jk} + e_{ijk} \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e_{ijk} \end{aligned}$$

(In der varianzanalytischen Darstellung, die in den jeweils oberen Zeilen der Gleichungen (8.1) und (8.2) gewählt wurde, kennzeichnen die  $\beta_k$  den Populationseffekt zu Lasten der Treatment-Modalität  $B_k$ , also  $\beta_k = \mu_k - \mu$ ; in den jeweils unteren Zeilen derselben Formeln bezeichnen die  $\beta_k$  die Populations-Regressions- bzw. Gewichtungskoeffizienten (vgl. detaillierter etwa Fennessey, 1968; Searle, 1971 a; Wottawa, 1974; Gaensslen & Schubö, 1976; Moosbrugger, 1978; Henning, 1978; Henning & Muthig, 1979).)

Die Regressionskoeffizienten  $\beta_k$  treten in den drei angegebenen grundlegenden Gleichungen stets nur in der 1. Potenz auf, weswegen man vom „Allgemeinen Linearen Modell“ spricht. Die Größen  $X_k$  können dabei in jeder beliebigen Potenz oder multiplikativen Verknüpfung auftreten; hierdurch wird die Form der sog. „Response Surface“ (auch: „Response Curve“) bestimmt - siehe dazu Box (1968), Bliss (1970, 427-433), John (1971, Kap. 10), Snedecor & Cochran (1972, 346-358) und John & Quenouille (1977, Kap. 9).

Die (wichtige) Unterscheidung zwischen der Modellform, die die Linearität der  $Y_{ik}$  als Funktion der unbekannten Regressionsgewichte  $\beta_k$  postuliert (Mendenhall, 1968, 55), und der Bedeutung der & für die Response Curve resp. Surface wird nicht immer beachtet, wie der Artikel von Brown (1975) zeigt; vgl. zu dieser Unterscheidung im einzelnen Mendenhall (1968, 51-55) und Henning (1978).

Die gebräuchlichen additiven Modellgleichungen repräsentieren dabei das einfachste Prinzip, nach dem man sich empirische Scores  $Y_{ik}$  zusammengesetzt denken kann. Es muß sich dabei nicht immer auch gleichzeitig um das

beste oder auch nur um ein adäquates Prinzip handeln, wie bereits Fisher & MacKenzie (1923, 315) ausgeführt haben. über andere, allerdings kaum gebräuchliche Möglichkeiten informieren etwa Searle (1971 a, 75 f.), Glass, Peckham & Sanders (1972, 240f.) sowie Namboodiri, Carter & Blalock (1975, 279f.).

Die Angemessenheit dieser Art von Modellgleichung ist im konkreten Einzelfall ebenso wenig empirisch prüfbar wie die entsprechende Modellannahme im Rahmen der klassischen Testtheorie (vgl. dazu Lord & Novick, 1968; Fischer, 1974; Kranz, 1979; Wottawa, 1980).

- (2) In mehrfaktoriellen Plänen spricht man daneben dann häufig von „Additivität“, wenn der Term  $(\alpha\beta)_{jk}$  oder  $\beta_3 X_1 X_2$  aus Formel (8.2) sowie alle Interaktionsterme höherer Ordnung gleich Null sind (vgl. Elston, 1961; Namboodiri, Carter & Blalock, 1975, 279-285). In bestimmten Versuchsanordnungen ist die Annahme, daß alle Interaktionsterme in den Populationen gleich Null sind, notwendig, um die interessierenden experimentellen Effekte auf statistische Signifikanz prüfen zu können (vgl. dazu neben Abschnitt 8.4.3 auch Wilk & Kempthorne, 1957; Scheffé, 1959, Kap. 5; Winer, 1971, 398f., 696f.; darüber hinaus weitere Lehrbücher der Versuchsplanung und -auswertung).

Diese Art der Additivität ist häufig einer statistischen Prüfung zugänglich. In den Fällen, in denen eine von der Interaktion unabhängige Fehler- und Prüfvarianz bestimmt werden kann (bspw. im Zufallsgruppenversuchsplan; vgl. dazu Abschnitt 3.3.2), erfolgt die Prüfung der Interaktion (oder Additivität) auf statistische Signifikanz unter Verwendung eben dieser Fehlervarianz. Für einige der Versuchspläne, in denen die die Interaktion enthaltende Residualvarianz die Prüfgröße darstellt, sind spezielle Testverfahren zur Prüfung der Additivitätsannahme entwickelt worden (Tukey, 1949; Johnson & Graybill, 1972). Die Leistungsfähigkeit dieser Tests ist u.a. von Hegemann & Johnson (1976) untersucht worden; Hinweise auf die Anwendungsbereiche finden sich etwa in den Arbeiten von Scheffé (1959, 130-134), Bliss (1967, 324-330, 451-465), Winer (1971, 394-397, 473-478), Bortz (1979, 398-400).

Ergibt die Anwendung eines dieser Testverfahren ein signifikantes Resultat, das für das Vorliegen von Nicht-Additivität oder das Vorhandensein einer Interaktion spricht, wird eine spezifische Voraussetzung für die Signifikanzprüfung im Rahmen der o. gen. Pläne über die F-Verteilung verletzt.

Als Ausweg wird in diesem Fall von zahlreichen (Lehrbuch-)Autoren empfohlen, die numerischen Werte der nicht-linear derart zu transformieren, daß die Wechselwirkung (Interaktion oder Nicht-Additivität) eliminiert wird - siehe z.B. Edwards (1971, 191f.) und Winer (1971, 398).

Dieser Empfehlung können wir uns nicht anschließen - zur Begründung siehe Abschnitt 8.2.4.2 -, sondern halten in der genannten Situation alternative Auswerteverfahren für eher angemessen - vgl. dazu Abschnitt 7.5.3.

### 8.2.3 Normalverteilung der Modellresiduen (Fehler)

In der Varianz- und Regressionsanalyse wird unter „Fehler“ derjenige Anteil am Score einer Vp verstanden, der in beliebiger Richtung vom Mittelwert der Gruppe abweicht, oder: der übrigbleibt, wenn den Daten eines Treatments ein bedingungsspezifischer Wert („fit“) angepaßt worden ist. Formal kann man für den einfachsten Fall etwa schreiben:

$$(8.3) \quad \text{Fehler} = \text{Residuum} = \text{beobachteter Wert} - \text{angepaßter Wert.}$$

Inhaltlich wird unter diese Fehler alles das subsumiert, was nicht auf den Einfluß der UV(n) zurückgeführt werden kann, also insbesondere individuelle Unterschiede (Persönlichkeitsmerkmale, -variablen; vgl. Herrmann, 1973, 1976) und ggf. Meßfehler (vgl. dazu Cochran, 1968b, 1970; Lord & Novick, 1968); es sei ergänzend nur angemerkt, daß gerade diese sog. Fehler unter anderen Fragestellungen der Gegenstand des Forschungsinteresses sind (siehe Herrmann, 1976). Wir gehen auf die eher inhaltlichen Aspekte der Fehlerterme nicht weiter ein, sondern verweisen auf die speziellen Arbeiten von Underwood (1957, Kap. 4 und 5), Cox (1961), Lord & Novick (1968), Cochran (1968b), Elashoff (1968), Mosteller (1968), Bredenkamp (1969a), Namboodiri, Carter & Blalock (1975, Kap. 12 und 13) sowie Henning & Muthig (1979, e.g. 35-37, 105-107).

Die Normalverteilung der individuellen Fehlerterme  $e_i$  in den einzelnen Populationen wird im Grunde vorausgesetzt, weil die Klasse von „normal“ genannten mathematischen Funktionen Eigenschaften aufweist, die zu Ableitungen von „bestechender Eleganz“ (Menges, 1972, 248) und Einfachheit führen; zu weiteren Gründen siehe Menges (a. a.o.). Bspw. sind Mittelwert  $\mu$  und Varianz  $\sigma^2$  dieser Verteilung stochastisch unabhängig voneinander (Geary, 1936), weswegen sich die Ableitung der Stichprobenverteilung des Mittelwerts  $M$  und der Varianz  $S^2$  vglw. problemlos gestaltet (vgl. Hays, 1977, Kap. 8 und 11).

Die bekannteren der zur Prüfung der Normalitätsannahme entwickelten Tests lassen sich grob in zwei Gruppen einteilen (vgl. dazu Rasch, Enderlein & Herrendörfer, 1973, Kap. 6), nämlich in Tests für die Anpassungsgüte (dazu u.a. Hays, 1977, Kap. 17; Bortz, 1979, 191-194) und in Tests für die Kennwerte „Schiefe“ („skewness“) und „Exzeß“ („kurtosis“) (dazu u.a. Gebhardt, 1966; Schmidtke & Jäger, 1976).

Die angesprochenen und verschiedene weitere, in der Literatur empfohlene Testverfahren sind in Simulationsstudien etwa von Shapiro, Wilk & Chen (1968) sowie von Saniga & Miles (1979) untersucht worden - auf diese Arbeiten und die dort enthaltenen Literaturangaben sei verwiesen.

Wir gehen auf diese Verfahren zur Prüfung der Normalverteilungsannahme nicht im einzelnen ein, da ihre routinemäßige Anwendung aus (mindestens) zwei Gründen kontraindiziert scheint.

Zum einen sind die entsprechenden Tests meist selbst abhängig von bestimmten Voraussetzungen oder Annahmen (vgl. etwa Box, 1953; Scheffé, 1959, 362; Gaensslen & Schubö, 1976, 58f.).

Zum anderen muß in fast jedem Fall mit einem signifikanten Resultat gerechnet werden, das für bedeutsame Abweichungen der Modellresiduen von der Normalität spricht, weil es nur wenige Gründe zu der Vermutung gibt, daß die Fehler sich *exakt* nach der „normal“ genannten Funktion verteilen: „Normality is a myth: there never was, and never will be, a normal distribution.“ (Geary, 1947, 241)

Wie Menges (1972, 249) ausführt, trifft diese Feststellung nicht notwendigerweise den „Kern der Dinge“: „... denn nicht nach der Existenz der Normalverteilung ist sinnvoll zu fragen, sondern nach dem Typ von empirischer Situation, für die das Modell, das sie repräsentiert, eine gute Approximation liefert.“ Demnach ist es zweckmäßig, die übliche Frage „Are normal theory ANOVA assumptions met?“ durch eine andere zu ersetzen, nämlich „How important are the inevitable violations of normal theory ANOVA assumptions?“ (Glass, Peckham & Sanders, 1972, 237).

Die auf eine „Ja-Nein-Antwort“ abzielende Frage wird also durch eine ersetzt, die nach der (qualitativ stuftbaren) „Robustheit“ der gebräuchlichen parametrischen Testverfahren fragt. Der Fachausdruck „Robustheit“ geht auf Box (1953, 318) zurück und bezeichnet die „Unempfindlichkeit (eines Testverfahrens) gegenüber Abweichungen von den postulierten Modellannahmen“ (vgl. Büning & Trenkler, 1978, 296).

In Abschnitt 7.5.2 hatten wir angesprochen, daß die Normalverteilung der Fehler eine der Voraussetzungen zur Ableitung der Stichprobenverteilungen von  $t$ ,  $\chi^2$  und  $F$  darstellt. Ist diese Voraussetzung nicht erfüllt, resultieren Verteilungen dieser Teststatistiken, die von den tabellierten zentralen und nicht-zentralen Verteilungen abweichen. Dies bedeutet, daß die tatsächlichen Wahrscheinlichkeiten für bestimmte Resultatsklassen nicht mehr mit den in den Tabellen angegebenen übereinstimmen, m.a. W. weichen die tatsächlichen Werte für  $\alpha$  und  $\beta$  von den festgesetzten nominellen Werten ab. Ein Test ist demnach robust, wenn unter Verletzung einer oder mehrerer Voraussetzungen

die tatsächlichen Werte für  $\alpha$  und  $\beta$  „nicht wesentlich“ von den nominellen abweichen - vgl. zum Versuch der quantitativen Definition von Robustheit die instruktive Arbeit von Bradley (1978); zum Hinweis auf einige Probleme beim ausschließlich qualitativen Robustheitskonzept siehe u.a. Büning & Trenkler (1978, 296f.) und Hager, Lübbecke & Hübner (im Druck).

Zur Beantwortung der Frage nach der Robustheit der klassischen parametrischen Teststatistiken  $t$ ,  $\chi^2$  und  $F$  unter Abweichung der Verteilung der Modellresiduen von der Normalität sind zahlreiche Untersuchungen durchgeführt worden (vgl. bereits Pearson, 1931; zusammenfassend Cochran, 1947; Lindquist, 1953, 78-90; Gaito, 1959b; Scheffé, 1959, Kap. 10; Bradley, 1968, 24-43; Glass, Peckham & Sanders, 1972; darüber hinaus im einzelnen u.a. Tikau, 1971; Kemp & Conover, 1973; Bevan, Benton & Myers, 1974; Feir-Walsh & Toothaker, 1974; Havlicek & Peterson, 1974; Lee, Desu & Gehan, 1975; Pearson & Please, 1975; Bowman, Beauchamp & Shenton, 1977; Posten, 1978; Trachtman, Giambalvo & Dippner, 1978; Bradley, 1980a, b, c; Blair, Higgins & Smitley, 1980).

Diese Studien an einfachen univariaten Versuchsplänen enthalten eine Fülle von Hinweisen darauf, daß die o.a. parametrischen Teststatistiken über einen weiten Bereich der Abweichungen von der Normalität der Fehlerverteilungen in den untersuchten Populationen robust sind, sofern sie zur Prüfung von Hypothesen über Mittelwerte oder quadrierte Korrelationsquadrate herangezogen werden (Modell der fixierten Effekte). Bei diesen Tests kann in aller Regel davon ausgegangen werden, daß die tatsächlichen Raten für Fehler 1. und 2. Art *nicht nennenswert* von den a priori spezifizierten nominellen Werten abweichen. Allerdings ist eine Quantifikation dieser Aussage etwa sensu Bradley (1978) u.E. derzeit noch nicht möglich, da die Befunde im Detail noch kein einheitliches Bild ergeben (siehe auch Büning & Trenkler, 1978, 296).

Im einzelnen ist mit *Überschreitungen* der nominellen Fehlerwahrscheinlichkeiten bei „breitgipfligen“ („platykurtischen“) Verteilungen zu rechnen, bei denen sich die Daten in den beiden Extrembereichen häufen; sehr ausgeprägt sind diese Abweichungen bei Stichproben aus L-förmigen Populationen, wie sie anscheinend Reaktionszeiten oft zugrundeliegen - siehe hierzu ausführlich Bradley (1968, 1977, 1978, 1980a, b, c).

Mit *Unterschreitungen* der nominellen Fehlerwahrscheinlichkeiten muß häufig bei „spitzgipfligen“ („leptokurtischen“) Verteilungen gerechnet werden. Abweichungen von der Symmetrie der Verteilungen wirken sich i. a. vornehmlich bei einseitigen Signifikanztests aus.

Komplexe Versuchspläne sind bislang nicht in der gleichen Ausführlichkeit und Systematik untersucht worden wie einfache Designs, weswegen die Robustheit von  $F$  und  $t$  für diese Fälle noch nicht beurteilbar ist.

Im Rahmen des varianzanalytischen Modells der zufälligen Effekte wird die Annahme normal-verteilter Treatment-Effekte gemacht, und der F-Test prüft eine Null-Hypothese bzgl. der Varianz dieser Effekte. Bei diesem Test haben insbesondere Abweichungen von der die Treatment-Effekte betreffenden Normalitätsannahme schwerwiegende Konsequenzen bzgl. der tatsächlichen Fehlerwahrscheinlichkeiten - vgl. hierzu etwa Lehmann (1968, 45), Glass & Stanley (1970, 462) sowie Hays (1977, 540f.).

Üblicherweise wird dann, wenn vermutet wird oder bekannt ist, daß die Fehler  $e_i$  in den Populationen nicht normal-verteilt sind, alternativ eine der beiden folgenden Empfehlungen gegeben:

1. Anwendung von nicht-parametrischen Verfahren.

Im Anschluß an unsere Erörterungen zu diesen Verfahren im Abschnitt 7.5.3.2 schließen wir uns dieser Empfehlung unter der Voraussetzung an, daß die Abweichungen von der Normalität sehr ausgeprägt sind oder aber daß außer der Normalitätsannahme noch mindestens eine weitere Voraussetzung nicht erfüllt ist.

2. Anwendung von Transformationen, die die Stichprobendaten „normalisieren“. Dieser Empfehlung können wir uns nicht anschließen - zur Begründung siehe Abschnitt 8.2.4.2.

### *8.2.4 Homogenität der Fehlervarianzen in den Populationen*

Eine weitere Voraussetzung für die valide Anwendung von F-Tests besteht in der Annahme der Varianzhomogenität, die besagt, daß die Modellresiduen in allen untersuchten Populationen gleiche (homogene) Varianzen aufweisen müssen.

Die Auswirkungen von Verletzungen dieser Voraussetzung auf das Verhalten der parametrischen Teststatistiken ist ebenfalls sehr ausführlich untersucht worden; entsprechende Zusammenfassungen der theoretischen und empirischen Resultate geben Cochran (1947), Scheffé (1959, Kap. 10), Lindquist (1953, 78-90), Box (1954a, b), Glass, Peckham & Sanders (1972); neuere Untersuchungen sind durchgeführt worden von Brown & Forsythe (1974a), Kohr & Garnes (1974), Feir-Walsh & Toothaker (1974), Ekbohm (1976), Keselman, Rogan & Feir-Walsh (1977), Havlicek & Peterson (1974), Howell & Garnes (1973), Rogan & Keselman (1977) und Hager et al. (im Druck).

Diesen Studien läßt sich entnehmen, daß in den meisten Fällen Varianzheterogenität tolerabel ist hinsichtlich der interessierenden tatsächlichen Werte für die Fehlerwahrscheinlichkeiten 1. und 2. Art, sofern die Stichprobengrößen für alle Bedingungen gleich sind. Nennenswerte Ausnahmen von dieser allgemeinen Aussage sind nur unter sehr extremen Bedingungen zu erwarten (vgl. Box, 1954a; Scheffé, 1959, 340; Rogan & Keselman, 1977).

Sind dagegen die Fehlervarianzen wie auch die Stichprobenumfänge ungleich, ist generell mit stärkeren Abweichungen der realen von den nominellen Fehlerwahrscheinlichkeiten zu rechnen (vgl. die o. a. Literatur).

Liegt dieser Fall vor, empfiehlt sich die Verwendung alternativer Testverfahren, etwa nicht-parametrischer (Abschnitt 7.5.3) oder spezieller parametrischer, auf die wir anschließend kurz eingehen werden. Zunächst wollen wir jedoch die Frage zu beantworten suchen, wie Varianzheterogenität festgestellt werden kann und ob dies sinnvoll ist.

Zur Prüfung einer Hypothese bzgl. der Gleichheit von Varianzen sind spezielle Testverfahren u.a. von Bartlett (1937), Hartley (1950) und Cochran (1951) entwickelt worden. Diese Verfahren sind jedoch ihrerseits empfindlich gegenüber Verletzungen der Normalverteilungsannahme - siehe dazu Pearson (1931), Box (1953), Scheffé (1959, 83-87), Levene (1960), Overall & Woodward (1974), Levy (1975a, 1978a) sowie den vorangegangenen Abschnitt.

Diese Tatsache hat zu einer Fülle von Versuchen geführt, Homogenitätsprüfungen für Varianzen zu entwickeln, die robust sind gegenüber Abweichungen von der Normalität (e.g. Box, 1953; Scheffe, 1959, 83; Levene, 1960; Glass, 1966; Leslie & Brown, 1966; Miller, 1968; Gartside, 1972; Layard, 1973; Brown & Forsythe, 1974c; Talwar & Gentle, 1977; O'Brien, 1978, 1979).

Die verschiedentlich angestellten Simulationsuntersuchungen zu den Testverfahren (vgl. Pearson, 1966; Garnes, Winkler & Probert, 1972; Hall, 1972; Layard, 1972; Brown & Forsythe, 1974c; Levy, 1975c, d, 1978a; Martin, 1976; Church & Wike, 1976; Martin & Garnes, 1977; Talwar & Gentle, 1977; O'Brien, 1978; Samiuddin, Hanif & Asad, 1978; Garnes, Keselman & Clinch, 1979a, b) haben zu recht unterschiedlichen Empfehlungen geführt, welches der Verfahren im konkreten Fall anzuwenden sei, und zwar in Abhängigkeit von der untersuchten Verteilung der Rohwerte in den Populationen. Da diese dem E üblicherweise nicht bekannt ist, ist uns eine explizite Empfehlung für einen bestimmten Varianzhomogenitätstest an dieser Stelle nicht möglich.

Es kann lediglich festgestellt werden, daß die *routinemäßige* Verwendung des in vielen Lehrbüchern der Versuchsplanung und -auswertung empfohlenen Tests nach Bartlett (1937) sicherlich *nicht* indiziert ist. Im Grunde scheint uns die Empfehlung, überhaupt auf routinemäßige Tests zur Prüfung der Varianzhomogenität zu verzichten, sofern die Stichprobenumfänge gleich groß sind, am ehesten der Mehrzahl der Fälle angemessen zu sein - zur Begründung dieser Aussage siehe die Erörterungen und Befunde von Horsnell, 1953; Box, 1953; Scheffé, 1959, 340f.; Young & Veldman, 1963; Glass, Peckham & Sanders, 1972, 242-246; Howell & Garnes, 1973; Kohr & Garnes, 1974; Havlicek & Peterson, 1974.

Für den Fall, daß die Stichprobenumfänge ungleich sind und eine Prüfung der Varianzhomogenität notwendig scheint, mag man eines der Verfahren auswählen, die in den

größeren der genannten Studien wiedergegeben und untersucht worden sind, also etwa bei Gartside (1972), Garnes, Winkler & Probert (1972), Hall (1972), Brown & Forsythe (1974c), Church & Wike (1976), Martin & Garnes (1977), O'Brien (1978). Zur Verwendung dieser Tests im Falle eines mehrfaktoriellen (varianzanalytischen) Versuchsplanes findet man einige wesentliche Hinweise etwa bei Overall & Woodward (1974), O'Brien (1978, 1979) und Garnes, Keselman & Clinch (1979b); zur Frage der Anschlußtests nach einem signifikanten generellen oder „Overall“-Test und der damit verbundenen Probleme orientiere man sich bei Levy (1975a, b), Garnes (1978a, b) sowie bei Garnes, Keselman & Clinch (1979a).

Ergeben diese Prüfungen, daß Varianzheterogenität vorliegt, ist bei gleichzeitigem Vorliegen von ungleichen Stichprobenumfängen von der Verwendung der Teststatistiken  $F$  und  $t$  abzuraten.

#### 8.2.4.1 Zur Frage des Prüfverfahrens bei Varianzheterogenität

Das Problem, welches Verfahren im Falle ungleicher Varianzen bei der Prüfung von Hypothesen über Mittelwerte aus zwei normalverteilten Populationen angemessen ist, wurde erstmals von Behrens (1929) und wenig später von Fisher (1935) zu lösen versucht; man findet daher oft in diesem Zusammenhang die Bezeichnung „Behrens-Fisher-Problem“. Die erste den Praktiker befriedigende Lösung dieses Problems wurde von Welch (1947, 1949, 1951), Aspin (1948, 1949) sowie James (1951) vorgeschlagen; weitere Lösungsansätze findet man bei Scheffé (1970) sowie Mehta & Srinivasan (1970) zusammengestellt und diskutiert. Das Welch-James-Verfahren kann deshalb als eine „befriedigende Lösung für den Praktiker“ apostrophiert werden, weil die von Aspin (1948, 1949), Trickett & Welch (1954), Trickett, Welch & James (1956) (vgl. auch Pearson & Hartley, 1962, Tab. 11) erarbeiteten Tabellen sehr genau über die  $t$ -Verteilung approximiert werden können, deren Freiheitsgrade „adjustiert“ worden sind - vgl. hierzu insbesondere Wang (1971); auch die Teststärke dieses Verfahrens kann im Vergleich zum  $F$ -Test als durchgängig gut bezeichnet werden, wie Golhar (1972) und Levy (1978 b) sowie Hager, Lübbecke & Hübner (im Druck) gezeigt haben.

Empirische Vergleiche des von Welch, James und Aspin empfohlenen Tests mit anderen zur Lösung des Behrens-Fisher-Problems vorgeschlagenen Verfahren mittels Simulationsstudien sind u.a. von Brown & Forsythe (1974b), Kohr & Garnes (1974), Ekbohm (1976), Levy (1978c, d), Keselman, Garnes & Rogan (1979a) angestellt worden. In diesen Studien konnte sich der hier empfohlene Test durchaus bewähren.

Um so unverständlicher ist es daher, daß er nicht wesentlich stärker in Standardlehrbüchern berücksichtigt wird; die für Psychologen wohl besten Darstellungen des Testverfahrens nach Welch, James und Aspin finden sich in Li (1964, 435-438), Winer (1971, 41-44), Rasch, Enderlein & Herrendörfer (1973, 80), Clauß & Ebner (1978, 213f.) sowie Pfanzagl (1978, 216f.).

Wichtig ist die abschließende Bemerkung, daß das empfohlene Verfahren auch für die Prüfung von mehr als zwei Mittelwerten, also auch für varianzanalytische Hypothesen, Verwendung finden kann (James, 1951; Welch, 1951; Brown & Forsythe, 1974a).



Neben dem parametrischen Test nach Welch et al. kann in den Fällen, in denen Varianzheterogenität gepaart mit ungleichen Stichprobengrößen auftritt, selbstverständlich ein nicht-parametrisches Verfahren zur Anwendung kommen. Darüber hinaus ist auf diese Gruppe von Verfahren zurückzugreifen, wenn mehr als eine der Voraussetzungen verletzt ist, weil in diesen Fällen starke Abweichungen von den nominellen Fehlerwahrscheinlichkeiten resultieren. Zum Vergleich der parametrischen Teststatistiken mit ihren nicht-parametrischen Analoga orientierte man sich im einzelnen bei Boneau (1962), Neave & Granger (1968), Kemp & Conover (1973), Feir-Walsh & Toothaker (1974), Lee, Desu & Gehan (1975), Keselman, Rogan & Feir-Walsh (1977), Blair, Higgins & Smitley (1980) und bei Hager et al. (im Druck).

#### 8.2.4.2 Zur Bedeutung von Transformationen

Zum Herstellen der Normalverteilung, zur Elimination von Nicht-Additivität (statistischen Interaktionen) und zur „Stabilisierung“ der Fehlervarianzen, d.h. Herbeiführung der Varianzhomogenität, werden nicht-lineare Transformationen der die Stichprobendaten generierenden Zufallsvariablen bereits seit geraumer Zeit vorgeschlagen, wie die Zusammenfassung von Bartlett (1947) ausweist - siehe ferner die einschlägigen Lehrbücher der Versuchsplanung und -auswertung.

Eine Systematisierung der verschiedenen möglichen Transformationen hat Lienert (1962) geleistet; weitere zusammenfassende Darstellungen finden sich bei Tukey (1949, 1957), Box & Cox (1964), Draper & Hunter (1969), Box & Tiao (1973, Kap. 10), Hoyle (1973), Schlesselman (1973), Smith (1976a) sowie Henning (1978).

Nur wenige Autoren nehmen eine kritische Haltung gegenüber Transformationen ein. Garnes & Lucas (1966) bspw. haben in Simulationsstudien gefunden, daß Transformationen zur Herstellung von normal verteilten Stichprobendaten nur selten zum gewünschten Ziel, nämlich validen Aussagen über Wahrscheinlichkeiten, führen (siehe auch Glass, Peckham & Sanders, 1972) und stellen fest: „In general, the use of a clearly interpretable scale of measurement certainly should be the dominant consideration“ (Garnes & Lucas, 1966, 326). Es ist insbesondere die Schwierigkeit der inhaltlichen Interpretation der (nicht-linear) transformierten Werte, die auch andere Autoren (e. g. Scheffé, 1959, 365-367; Digman, 1966, 475; Lindman, 1974, 35; Namboodiri, Carter & Blalock, 1975, 286) zu ähnlichen Empfehlungen gelangen lassen. Für uns ist ein anderer Grund von mindestens gleichrangiger Bedeutung, wenn wir dazu raten, auf nicht-lineare Transformationen zu verzichten: Ist die theoretische AV ein metrischer Begriff und muß für die empirische AV deshalb Intervallskalenniveau angestrebt werden, sind ausschließlich lineare Transformationen zulässig. Hat die empirische AV nur Ordinalskalenniveau, sind die hier betrachteten parametrischen Testverfahren nicht sinnvoll anwendbar, weil sich die Ergebnisse dieser Tests unter erlaubten und willkürlich möglichen monotonen Transformationen der empirischen AV ändern können (vgl. hierzu Abschnitt 2.4).

Eine ganz ähnliche Argumentation findet sich auch bei Henning (1978), der eine der wenigen Arbeiten verfaßt hat, die die Transformationen sowohl unter mathematischen als auch unter meßtheoretischen Aspekten untersuchen; seine zusammenfassende und

vertiefende Erörterung der bei Datentransformationen möglichen Probleme sei dem Interessierten empfohlen. Die Darstellung der wesentlichen Gedanken dieser Arbeit findet sich auch in Henning & Muthig (1979, 205-213).

### 8.2.5 Unabhängigkeit der Fehlerterme

Die statistischen Fehler oder Residuen als nicht auf die systematische Bedingungsvariation zurückführbare Komponenten von empirischen Scores können auf verschiedene Arten voneinander und von anderen Komponenten abhängig sein.

Bei den folgenden Darstellungen bezeichnen wir mit  $e_s$  und  $e_{s'}$  ( $s \neq s'$ ) zwei unterschiedliche Fehlerkomponenten und mit  $m$  und  $m'$  ( $m \neq m'$ ) zwei verschiedene Treatmentbedingungen.

- (1) Innerhalb eines Treatments  $m$  soll die Korrelation  $R$  zwischen allen möglichen Paaren von Fehlertermen gleich Null sein:  $R(e_{sm} e_{s'm}) = 0$ ; zur Berechnung siehe etwa Snedecor & Cochran (1972, 294f.).

Korrelierte Fehler *innerhalb* einer experimentellen Gruppe können etwa entstehen, wenn bei der Zuweisung der  $V_{pn}$  zu den Treatments wissentlich oder unwissentlich nach einer bestimmten Systematik verfahren wird, so daß sich die  $V_{pn}$  innerhalb einer Experimentalgruppe sehr ähnlich verhalten.

- (2) Zwischen je zwei experimentellen Bedingungen soll die (Inter-)Korrelation zwischen allen möglichen Paaren von Fehlern gleich Null sein:  
 $R(e_{sm} e_{s'm'}) = 0$ .

In der Regel führen (unkontrollierte oder Stör-)Bedingungen, die gemeinsam mit einer bestimmten Realisation der UV auftreten, zu einer Konfundierung von Variablen, die auch die Unabhängigkeit der Fehler zwischen zwei beliebigen Treatments beeinträchtigt. Da hierbei der Einfluß der experimentell untersuchten UV nicht mehr von den Störbedingungen getrennt werden kann, liegt auch eine Verletzung der internen Validität vor.

Wird bei Vorliegen von korrelierten Fehlern der unter (1) und (2) beschriebenen Art ein F-Test durchgeführt, so sind im Falle des Vorliegens von positiven Korrelationen zu viele fälschlich signifikante Resultate zu erwarten, weil der Erwartungswert der Teststatistik  $F$  unter Gültigkeit der  $H_0$  größer ist als  $df_N / (df_N - 2)$ . Im seltener vorliegenden Fall von negativen Korrelationen ist dagegen mit einer ungerechtfertigt erhöhten Anzahl von nicht-signifikanten Ergebnissen in F-Tests zu rechnen, weil der Erwartungswert von  $F$  unter  $H_0$  kleiner als 1 werden kann. Nähere Einzelheiten hierzu finden sich u.a. in Cochran (1947), Glass & Stanley (1970), Snedecor & Cochran (1972), Keppel (1973), Lissitz & Chardos (1975) sowie Hays (1977).

Insgesamt ist festzuhalten, daß der F-Test nicht robust ist, wenn die Annahme der unabhängigen Fehlerterme nicht erfüllt ist. Daher ist es unabdingbar, die beiden unter (1) und (2) angesprochenen Abhängigkeiten zwischen Fehlertermen zu vermeiden. Dies ist zu erreichen, indem man die Vpn zufällig auf die Untersuchungsbedingungen verteilt - siehe Box (1954b), Bliss (1967, 340), Bredenkamp (1969a, 338), Öchran & Bliss (1970, 45-47). Ferner muß der E dafür Sorge tragen, daß die Werte der AV innerhalb jeder und zwischen allen experimentellen Gruppen unabhängig voneinander erhoben werden (können). Hieraus folgt nicht, daß keine Gruppenexperimente mehr durchgeführt werden sollten; die Forderung nach Unabhängigkeit bezieht sich auf die Werte der AV, die der statistischen Analyse unterzogen werden - vgl. zur weiteren Erörterung des Unterschiedes zwischen „experimentellen Einheiten“ und „Einheiten der statistischen Analyse“ etwa Glass & Stanley (1970, 501-508) sowie Abschnitt 3.3.2.

Die in diesem Abschnitt erfolgte sehr kurze Darstellung einiger der wesentlichsten möglichen Abhängigkeitsbeziehungen zwischen den verschiedenen Komponenten eines empirischen Scores, die meist nur dann erfaßbar sind, wenn man den Daten (a priori) ein entsprechendes statistisches Modell anpaßt, kann durch die Lektüre insbesondere von Hays (1977, 467, 481-483, 502, 528f., 535f., 540-543, 553, 568-574) sowie darüber hinaus etwa von Scheffé (1959, 333-339), Keppel (1973, 76, 199f., 462-467) und Bortz (1979, 344-347) wesentlich vertieft werden.

#### 8.2.5.1 Zur Residuenanalyse; Ausreißerwerte

Bei genauerer Betrachtung wird offensichtlich, daß die Mehrheit der für die valide Anwendung der parametrischen Signifikanztests notwendigen Voraussetzungen sich auf die (theoretische) Verteilung der Fehler oder Residuen  $e_{im}$  beziehen. Diese stellen den individuellen Anteil an den Daten dar, der nicht auf systematische Bedingungseinflüsse zurückführbar ist.

Wegen der daraus sich ergebenden Bedeutung, die den Residuen im Rahmen des Konzepts der statistischen Validität zukommt, ist eine Inspektion dieser Fehler nach der Datenerhebung durchaus zu empfehlen; zur weiteren Begründung dieser Forderung vgl. etwa das Beispiel von Broekman (1973) und das Zahlenmaterial in Hampel (1980, 9). Spezielle Techniken zur Analyse der Fehlerterme sind von Anscombe & Tukey (1963), Draper & Smith (1966, Kap. 3), Wooding (1969), Behnken & Draper (1972) und Tukey (1977) vorgestellt worden. Unabhängig davon, ob man sich dieser teilweise sehr elaborierten Verfahren bedienen will, ist eine graphische Veranschaulichung der Rohwerte oder aber der Fehler als Bestandteil der erhobenen Rohwerte in der Regel unerläßlich - einen Überblick über graphische Darstellungsformen und Techniken findet man bei Wainer & Thissen (1981).

Durch Inspektion dieser Graphen etc. erhält man einen ersten Eindruck über die Verteilung der Fehler (in der Stichprobe). Insbesondere kann man hierbei sog. „Ausreißer-“ oder „Extremwerte“, die „Outliers“, entdecken, also einzelne Roh- oder Fehlerwerte, die sich durch einen ziemlichen Abstand von der Masse der Daten auszeichnen, d.h. die im Vergleich zu den übrigen Werten entweder ungewöhnlich niedrig oder hoch sind.

Für die Entstehung derartiger Outliers sind nun mindestens drei Gründe denkbar (vgl. etwa Cochran & Bliss, 1970, 47; Hampel, 1980):

- (1) Es wurden Fehler bei der Datenerhebung gemacht, etwa falsches Registrieren, Ablesen oder Übertragen von Werten.

Sofern der richtige Wert zu ermitteln ist, sollte dieser den falschen Wert ersetzen. Im anderen Fall empfiehlt es sich, den falschen Wert ersatzlos zu streichen und als „missing value“ zu behandeln - ggf. ist dann in den restlichen experimentellen Bedingungen ebenfalls je ein Wert (zufällig) zu eliminieren (siehe dazu Abschnitt 10.4.2).

- (2) Der oder die Extremwert(e) deutet (deuten) darauf hin, daß die Populationsverteilung der Fehler von einer Normalverteilung stark abweicht.
- (3) Die Fehlerwerte folgen zwar einer Normalverteilung, aber die festgestellte Ausreißerwert repräsentiert das sog. „seltene Ereignis“, von dem der E stets annimmt, daß es ihm nicht widerfährt (vgl. Fisher, 1951, 14; Neyman, 1952, 43; Haagen & Seifert, 1979, 185).

Derartige Ausreißerwerte können die numerische Größe der gebräuchlichen parametrischen Statistiken „Mittelwert  $M$ “ und „Varianz  $S^2$ “ sehr stark beeinflussen bzw. verzerren; diese Statistiken sind nicht „resistent“ (oft auch: „robust“) gegenüber Ausreißern - vgl. zu Beispielen und weiteren Einzelheiten etwa Mosteller & Tukey (1977, 203-212).

Die Tatsache der unzulänglichen Resistenz der „klassischen“ parametrischen Statistiken gegenüber Ausreißern hat zu unterschiedlichen Empfehlungen geführt, wie man sich angesichts von Extremwerten verhalten sollte.

- (1) Man verzichtet auf die parametrischen Statistiken  $M$ ,  $S^2$  und  $r$  und ersetzt sie durch resistente analoge Kenngrößen wie etwa den Median  $M_d$  und Rangkorrelationskoeffizienten oder aber durch spezielle, besonders in jüngster Zeit vorgeschlagene Statistiken, über die im einzelnen etwa Andrews et al. (1972), Wainer (1976), Wainer & Thissen (1976), Tukey (1977) sowie Mosteller & Tukey (1977) informieren. über diese analogen Statistiken können häufig die üblichen Signifikanztests, teilweise leicht modifiziert, durchgeführt werden (vgl. u.a. Schrader & McKean, 1977).
- (2) Man verwendet zwar die üblichen Statistiken, modifiziert oder eliminiert jedoch die Ausreißerwerte.

Mittels spezieller Testverfahren kann bspw. beurteilt werden, ob die Ausreißerwerte einer Normalverteilung entstammen und daher nur ein zufällig extremes Resultat darstellen - eine zusammenfassende Darstellung dieser Verfahren findet sich bei Barnett & Lewis (1978); neuere Tests wurden von Tiku (1975) und Hawkins (1979) vorgestellt. Die Tests entdecken erst vergleichsweise extreme Ausreißerwerte.

Bei Vorliegen von zufälligen, einer Normalverteilung entstammenden extremen Residuen werden unterschiedliche Strategien empfohlen, die darauf abzielen, den Einfluß dieser Werte auf die Größe der zu berechnenden Statistiken zu minimieren.

Die einfachste Lösung stellt das sog. „Stutzen“ oder „Trimming“ dar. Hierbei werden die extremen Werte bei der Analyse unberücksichtigt gelassen. Eine andere Lösung, die sog. „Winsorization“, ersetzt die Outliers nach einem bestimmten Modus durch weniger extreme Daten. Einzelheiten zu beiden Ansätzen sind den Publikationen von Tukey & McLaughlin (1963), Dixon & Tukey (1968), Andrews et al. (1972), Wainer (1976), Mosteller & Tukey (1977) und Hampel (1980) zu entnehmen. All die vorgeschlagenen Maßnahmen, deren grundsätzliche Rechtfertigung und Validität durch eine großangelegte Computer-Studie belegt wird (Andrews et al., 1972; ergänzend Wegman & Carroll, 1977), entheben den E jedoch nicht der Notwendigkeit, das Zustandekommen der Outliers besonders zu diskutieren: „Sometimes the outlier is providing information which other data points cannot due to the fact that it arises from a unusual combination of circumstances which may be of vital interest and requires further investigation rather than rejection.“ (Draper & Smith, 1966, 95)

Weitere Informationen zum Komplex der Residuenanalyse und der Resistenz von statistischen Kennwerten gegenüber Extremwerten können den zusammenfassenden Arbeiten etwa von Huber (1972), Hogg (1974, 1977, 1979), Bickel (1976), Wainer (1976) sowie Hampel (1980) entnommen werden. über eine spezielle Technik im Zusammenhang mit Varianzen, das „Jackknifing“, informieren zudem u.a. die Arbeiten von Miller (1968, 1974), Mosteller & Tukey (1968, 1977) sowie Gray & Schucany (1972).<sup>27)</sup>

U. E. sollte eine Sonderbehandlung von Extremwerten ausschließlich dann in Erwägung gezogen werden, wenn dadurch die Validität des Experiments zur Prüfung der interessierenden WH erhöht wird. Dies ist beispielsweise dann der Fall, wenn die Extremwerte auf fehlerhafte Datenerhebung beruhen - nach Hampel (1980, 9) eine der häufigsten Ursachen für Ausreißer - oder wenn vermutet werden muß, daß sie von Vpn stammen, die die Instruktionen nicht angemessen befolgt haben.

<sup>27)</sup> „Jackknifing“ wird üblicherweise auf Varianzen angewendet, ist jedoch nicht auf diese Anwendung beschränkt (siehe Gray & Schucany, 1972).

### 8.2.6 Problem der Zufallsstichproben

Eine der Voraussetzungen für parametrische Testverfahren besteht darin, daß die zu untersuchenden Daten eine (oder mehrere) Zufallsstichprobe(n) aus einer (oder mehreren) (in der Regel normalverteilten) Populationen darstellen. Nur dann lassen sich die Wahrscheinlichkeiten für bestimmte Realisationen von Statistiken angeben. Diese Angaben wiederum sind notwendig, um überhaupt statistische Tests im oben beschriebenen Sinne (vgl. Teil 7) durchführen zu können, bei denen die Wahrscheinlichkeiten für Fehlentscheidungen - und sei es auch nur innerhalb gewisser Grenzen - ange- und kontrollierbar sind.

Mit dem Begriff „Zufallsstichprobe“ wird eine bestimmte mathematische Modellvorstellung bezeichnet (s. Fisz, 1970, 394; Menges, 1972), die angemessen ist, wenn aus einer endlichen Grundgesamtheit eine Teilmenge so ausgewählt wird, daß jedes Element der Grundgesamtheit mit der gleichen Wahrscheinlichkeit „gezogen“ wird. In der Psychologie werden Untersuchungen aber in aller Regel mit den experimentellen Einheiten durchgeführt, die gerade „verfügbar“ sind, die in diesem Sinne also keine Zufallsstichprobe aus einer bestimmten existierenden Population sind. Man kann parametrische Tests aber auch unter diesen Umständen anwenden:

Viele parametrische Tests können als rechnerische Vereinfachungen entsprechender Randomisierungstests interpretiert werden (s. Abschn. 7.5.2). Bei Randomisierungstests wird die Verteilung der Teststatistik unter Gültigkeit der Nullhypothese aber abgeleitet, ohne daß Annahmen über Eigenschaften von Populationen gemacht werden. Deshalb können diese Tests unabhängig davon durchgeführt werden, aus welchen Populationen die beobachteten Daten stammen. Es ist dann aber auch möglich, jeden Datensatz als Zufallsstichprobe aus einer *hypothetischen* Population anzusehen (Alf & Abrahams, 1973), deren Verteilungsfunktion genau jene Eigenschaften hat, bei denen die Wahrscheinlichkeit maximiert wird, daß die empirischen *Daten* eine Zufallsstichprobe aus dieser Population sind. Die von Randomisierungstests geprüften Nullhypothesen sind Aussagen über diese hypothetischen Populationen, z.B. über die Gleichheit der Verteilungsfunktionen der Populationen, die den verschiedenen Gruppen von Daten (also etwa den experimentellen Bedingungen) entsprechen, oder über die Unabhängigkeit von Variablen in diesen Populationen (siehe Bradley, 1968, 68-73).

Insgesamt ist es zur Durchführung und Interpretation parametrischer Testverfahren also nicht notwendig, daß die beobachteten *experimentellen Einheiten* eine Zufallsstichprobe aus einer real existierenden oder inhaltlich definierbaren Population darstellen. Diese Frage der statistischen Validität ist zu unterscheiden von der Frage, an welchen experimentellen Einheiten eine Kausalhypothese

se zu prüfen ist, damit die entsprechende Untersuchung eine möglichst hohe Populationsvalidität hat. Nach den Ergebnissen des Abschnitts 4.1 sind aber auch dazu keine Zufallsstichproben notwendig.

### 8.3 Kumulierung der Wahrscheinlichkeiten für Fehler erster und zweiter Art

Führen wir *einen* Signifikanztest durch, ist die Wahrscheinlichkeit für eine fälschliche Ablehnung von  $H_0$  höchstens gleich  $\alpha$ . Führen wir *zwei* unabhängige Signifikanztests durch, und ist in beiden Fällen  $H_0$  richtig, so ist die Wahrscheinlichkeit, daß in beiden Fällen auch richtigerweise ein für  $H_0$  sprechendes Stichprobenergebnis eintritt, gleich  $(1 - \alpha)^2$ .

Die Komplementärwahrscheinlichkeit, daß einer oder beide Tests fälschlicherweise für  $H_1$  sprechen, ist damit gleich  $1 - (1 - \alpha)^2$ . Allgemein berechnet sich daher die Wahrscheinlichkeit, bei  $T$  unabhängigen Signifikanztests, bei denen jeweils die Nullhypothese zutrifft, *mindestens* ein für  $H_1$  sprechendes Stichprobenergebnis zu erhalten, zu:

$$(8.4) \quad P_\alpha = 1 - (1 - \alpha)^T$$

(Ryan, 1959, 1962; Hays, 1977, 611). Bei ungleichen  $\alpha$ -Niveaus für die einzelnen Tests ergibt sich die Gesamtwahrscheinlichkeit entsprechend aus

$$(8.5) \quad P_\alpha = 1 - (1 - \alpha_1) \cdot (1 - \alpha_2) \cdot \dots \cdot (1 - \alpha_T) = 1 - \prod_{t=1}^T (1 - \alpha_t)$$

(Miller, 1966, 8).

Sind die statistischen Tests nicht unabhängig, stellt  $P_\alpha$  zugleich eine gute Annäherung und eine *obere Schranke* für die gesuchte Wahrscheinlichkeit dar (Kimball, 1951; Miller, 1966, 101-102, 1977; Petrinovich & Hardyck, 1969).

Nach (8.4) steigt bei Durchführung mehrerer Signifikanztests die Wahrscheinlichkeit von mindestens einem zufällig signifikanten Ergebnis schnell an, bei einem  $\alpha = 0,05$  und den 7 Tests einer dreifachen Varianzanalyse kann beispielsweise  $P_\alpha$  schon 0,30 betragen, bei den 15 möglichen Tests einer vierfachen Varianzanalyse schon 0,54 (vgl. die Tabellen bei Jacobs, 1976). Deshalb sollten komplexe Versuchspläne, die mehrere Signifikanztests erforderlich machen, nur verwendet werden, wenn es zur Prüfung der betrachteten wissenschaftlichen Hypothese unumgänglich ist. Insbesondere sollten in Varianz- und Regressionsanalysen nur solche Faktoren berücksichtigt und nur solche (Haupt- und Interaktions-)Effekte geprüft werden, die tatsächlich von theore-

tischer oder praktischer Wichtigkeit sind (vgl. Rule, 1976). Hat man dennoch mehrere Signifikanztests durchzuführen, kann man den Kumulationseffekt dadurch ausgleichen, daß man bei T Tests die Irrtumswahrscheinlichkeit  $\alpha$  jedes Einzeltests so wählt, daß  $P_\alpha$  höchstens gleich einem festzusetzenden Grenzwert  $\alpha_G$  ist. Mit  $P_\alpha = \alpha_G$  ergibt sich aus (8.4) nämlich:

$$(8.6) \quad \alpha = 1 - \sqrt[T]{1 - \alpha_G}$$

(vgl. Garnes, 1971b). Der Wert  $\alpha_0$  gibt dabei die festgesetzte, auf das gesamte Experiment bezogene Wahrscheinlichkeit für mindestens einen Fehler 1. Art an. Der Einfachheit halber kann man auch die aus der *Bonferroni-Ungleichung* abgeleitete Regel

$$(8.7) \quad \alpha = \alpha_G / T$$

(Dunn, 1961; Dunnett, 1970; Garnes, 1971b) verwenden, die nur zu ganz geringfügig niedrigeren Werten führt (vgl. Lunney, 1969).<sup>28)</sup>

Ein Nachteil dieser Vorgehensweisen ist darin zu sehen, daß durch das Herabsetzen der Fehlerwahrscheinlichkeit  $\alpha$  für die einzelnen Tests bei gleichbleibender Zahl der Beobachtungen die Teststärke  $1 - \beta$  der Tests sinkt, bzw. daß mehr Beobachtungen notwendig sind, um eine gleichbleibende Power zu erzielen. Diese verminderte Effizienz stellt auch den entscheidenden Nachteil der Verfahren für multiple Mittelwertsvergleiche dar, die von vornherein von einem adjustierten Signifikanzniveau ausgehen, wie es etwa bei den Tests nach Tukey und Scheffé der Fall ist (vgl. Abschnitt 8.3.1).

Die dargestellten traditionellen Überlegungen zur Kumulierung von Fehlern 1. Art sind noch ergänzungsbedürftig, wenn wir wieder stärker die strenge Prüfung wissenschaftlicher Hypothesen als übergeordnetes Ziel aller unserer (experimentellen) Bemühungen berücksichtigen. Werden bei der empirischen Überprüfung einer Hypothese oder Theorie nicht nur eine, sondern mehrere statistische Hypothesen aus ihr abgeleitet, handelt es sich meist um Alternativhypothesen  $H_{1(1)}, H_{1(2)}, \dots, H_{1(T)}$  (vgl. die Punkte (6), (7) und (8) in Abschnitt 8.1.1). Wird die wissenschaftliche Hypothese nur dann als bewährt betrachtet, wenn alle implizierten Alternativhypothesen angenommen werden können, ist die Wahrscheinlichkeit, daß man diese Entscheidung fälschlicherweise trifft, stets höchstens gleich  $\alpha$ . Deshalb muß ein konventionell beispielsweise auf 0,05 festgesetzter Wert für  $\alpha$  nicht adjustiert werden (Westermann, im Druck, b; Hager & Westermann, im Druck, b).

---

<sup>28)</sup> Zur Wahl von  $\alpha$ , wenn die Wahrscheinlichkeit für mindestens  $r$  ( $r = 2, 3, \dots$ ) Fehler erster Art einen festen Wert  $\alpha_G$  nicht überschreiten soll, siehe Hsu (1978), Hurlburt & Spiegel (1976) und die Tabellen bei Feild & Armenakis (1974).



Auf der anderen Seite könnte man sich für eine Falsifikation der wissenschaftlichen Hypothese entscheiden, wenn mindestens eine der implizierten statistischen Hypothesen nicht zutrifft. Durch Übertragung der Ergebnisse aus den Überlegungen zur Kumulierung der Fehler erster Art wissen wir, daß bei Gültigkeit der statistischen Alternativhypothesen  $H_{1(1)}, H_{1(2)}, \dots, H_{1(T)}$  die Wahrscheinlichkeit für mindestens ein (fälschlich) für eine Null-Hypothese  $H_{0(t)}$  sprechendes empirisches Ergebnis bis auf

$$(8.8) \quad P_\beta = 1 - \prod_{t=1}^T (1 - \beta_t),$$

ansteigen kann.

Impliziert eine wissenschaftliche Hypothese mehrere statistische Alternativhypothesen, muß man daher zur Vermeidung ungerechtfertigter Falsifikationen die  $\beta$ -Wahrscheinlichkeit für jeden einzelnen Test möglichst klein wählen (Westermann, im Druck, a). Idealerweise sollte sie analog zu (8.7) durch

$$(8.9) \quad \beta = \beta_G/T$$

bestimmt werden.

Beispiel: Soll bei Durchführung von 3 Signifikanztests die Wahrscheinlichkeit von mindestens einer fälschlichen Annahme der Nullhypothese höchstens gleich 0,20 sein, muß die Zahl  $N$  der Beobachtungen so gewählt werden, daß jeder Signifikanztest eine Teststärke von  $1 - \beta = 1 - \beta_G/T = 1 - 0,20/3 = 0,93$  aufweist.

Diese Überlegungen hinsichtlich einer Herabsetzung von  $\alpha$  und  $\beta$  zum Ausgleich von Kumulierungseffekten können auf alle Entscheidungen zwischen zwei Alternativen  $A$  und nicht- $A$  verallgemeinert werden, die aufgrund der Ergebnisse mehrerer Signifikanztest getroffen werden: Wird eine Entscheidung (z.B. für  $A$ ) getroffen, wenn alle statistischen Hypothesen einer bestimmten Menge angenommen werden können, muß die Wahrscheinlichkeit für die fälschliche Annahme einer dieser Hypothesen nicht adjustiert werden; fällt die Entscheidung (für  $A$ ) aber schon, wenn (*mindestens*) eine der statistischen Hypothesen angenommen wird, sollte die Wahrscheinlichkeit für eine Fehlentscheidung auf die Gesamtmenge der statistischen Tests bezogen werden, für den einzelnen Test also herabgesetzt werden (vgl. Rule, 1976).

Diskutiert wird häufig die Frage, wie die „Familie“ von Tests zu spezifizieren ist, für die die Irrtumswahrscheinlichkeit insgesamt festgelegt wird (Miller, 1966, 31-35; Ryan, 1962; Kirk, 1968, 77f.; Dunnett, 1970, 100). Uns erscheint es ungerechtfertigt, generell alle Signifikanztests eines Experiments zusammenzufassen („experimentwise error rate“), da es z.B. recht willkürlich sein kann, ob eine Hypothese nun allein oder (in einem multifaktoriellen

Design) zusammen mit anderen überprüft wird. Die obigen Erörterungen bieten u.E. bessere Leitlinien: Für mehrere Signifikanztests ist genau dann insgesamt eine maximale Fehlerwahrscheinlichkeit festzulegen, wenn mit der Annahme einer einzigen der statistischen Hypothesen die Voraussetzung für eine bestimmte Entscheidung (z.B. über die Falsifikation einer wissenschaftlichen Hypothese oder die Einleitung einer praktischen Maßnahme) erfüllt ist.

Im folgenden wollen wir die Behandlung des Problems der kumulierten Fehlerwahrscheinlichkeiten für zwei besonders wichtige Hypothesenarten noch konkreter betrachten.

### 8.3.1 Multiple Mittelwertsvergleiche

Von multiplen Mittelwertsvergleichen wollen wir sprechen, wenn zugleich mehrere statistische Hypothesen über eine Menge von Mittelwerten geprüft werden. Dies kann in folgenden Situationen der Fall sein:

- (1) Die wissenschaftliche Hypothese impliziert die Null- oder Alternativhypothese einer Varianzanalyse (siehe Punkt (6) in Abschnitt 8.1.1). Die Annahme der statistischen  $H_1$  ist dann gleichbedeutend mit der Entscheidung, daß mindestens zwei der betrachteten Mittelwerte  $\mu_1, \dots, \mu_K$  ungleich sind. In diesem Fall kann es von Interesse sein zu prüfen, welche der Mittelwerte dies sind und/oder ob ganz bestimmte Mittelwerte ungleich sind.
- (2) Die psychologische Hypothese impliziert verschiedene statistische Hypothesen über eine Menge von Mittelwerten (vgl. Punkt (8) im Abschnitt 8.1.1), also z.B. die folgenden vier Null- bzw. Alternativhypothesen:

$$H_{0(1)}: \mu_1 = \mu_2$$

$$H_{0(2)}: \mu_3 = \mu_4$$

$$H_{0(3)}: \mu_4 = \mu_5$$

$$H_{1(4)}: (\mu_1, \mu_2) > (\mu_3, \mu_4, \mu_5)$$

Dabei bezeichnet z.B.  $(\mu_3, \mu_4, \mu_5)$  den Mittelwert  $\mu_{3,4,5}$  der aus den Populationen 3, 4 und 5 gebildeten Gesamtpopulation, der sich zu  $(1/3)(\mu_3 + \mu_4 + \mu_5)$  ergibt. Geprüft werden derartige Hypothesen über entsprechend konstruierte Linearkombinationen von Stichprobenmittelwerten („Methode der (orthogonalen) Kontraste“; siehe dazu u.a. Winer, 1971, 170-177; Myers, 1972, 352-362; Hays, 1977, 581-605).

Jeder Mittelwertshypothese entspricht eine Hypothese über einen Kontrast, d.h. eine Linearkombination von Mittelwerten, die unter den üblichen Annahmen des ALM (vgl. Abschnitt 8.2.1) über t-verteilte Teststatistiken geprüft

werden kann.<sup>29)</sup> Nach dem zu Beginn dieses Abschnitts Gesagten muß je nach Fragestellung  $\alpha$  und/oder  $\beta$  für die Einzeltests so gewählt (adjustiert) werden, daß die Entscheidung für oder gegen die wissenschaftliche Hypothese nicht allein aufgrund der erhöhten Wahrscheinlichkeit für mindestens einen Fehler 1. oder 2. Art getroffen wird. Wir wollen dann von adjustierten multiplen t-Tests sprechen.

Die Bestimmung kritischer t-Werte ist auch für „ungewöhnliche“ Signifikanzniveaus  $\alpha$  nicht schwierig. Sie erfolgt über die ausführlichen Tabellen der t-Verteilung bei Pearson & Hartley (1962, 132-134) oder über die speziell für diesen Zweck eingerichteten Tabellen von Dunn (1961; auch in Kirk, 1968, 551; und Marascuilo & McSweeney, 1977, 483) oder von Dayton & Schafer (1973).

Auch wenn aus der wissenschaftlichen Hypothese eine Menge von T Mittelwerts-hypothesen folgt, kann es sinnvoll sein, vor Durchführung der T Mittelwertsvergleiche als globalen oder „Overall-Test“ eine Varianzanalyse oder eine ihrer nicht-parametrischen Entsprechungen über alle Mittelwerte durchzuführen (vgl. Fisher, 1951, 196-204; Miller, 1966, 90; Dunnett, 1970; Carmer & Swanson, 1971, 1973; Cohen & Cohen, 1975, 162-165; Gaito, 1978; Keselman, Garnes & Rogan, 1979a, 1980; Swaminathan & deFriesse, 1979; Ryan, 1980). Dabei muß die Zahl der Beobachtungen so gewählt werden, daß  $\alpha$  und  $\beta$  auf kleine Werte fixiert werden können. Wird dann der Test der Null-Hypothese, daß alle Mittelwerte gleich sind, nicht signifikant, können auch die Null-Hypothesen aller interessierenden Mittelwerts-*kontraste* angenommen werden.

Impliziert also die wissenschaftliche Hypothese die Gültigkeit der Alternativ-hypothesen dieser Kontraste, kann sie damit bereits aufgrund des globalen Tests falsifiziert werden (vgl. im einzelnen Teil 11), ohne daß es zur Kumulierung von Fehlerwahrscheinlichkeiten kommen kann. Wird dagegen die  $H_1$  des Overall-Tests angenommen, kann mit der relativ klein angesetzten Irrtumswahrscheinlichkeit  $\alpha$  geschlossen werden, daß *mindestens* zwei Mittelwerte unterschiedlich sind. Ob dies allerdings auch die von der wissenschaftlichen Hypothese implizierten sind, muß anschließend über die multiplen Vergleiche geprüft werden. Analoges gilt, wenn aus der psychologischen Hypothese die Null-Hypothese einer Menge von Mittelwertsvergleichen folgt.

---

<sup>29)</sup> Zur nicht-parametrischen Prüfung von Kontrasten siehe vor allem Marascuilo & McSweeney (1967, 1977, S. 306-312, 362-371); ferner Keselman & Rogan (1977), Wike & Church (1977, 1978), Silverstein (1978), Church & Wike (1979, 1980), Levy (1979a, b) sowie Shuster & Boyert (1979).

Sind aus der wissenschaftlichen Hypothese nicht Hypothesen über Mittelwerte, sondern über andere Parameter abgeleitet worden (etwa über Varianzen, Korrelationen oder Porportionen), dann lassen sich diese ebenfalls über die hier beschriebene Technik der multiplen Vergleiche prüfen (siehe Garnes, 1978a, b); das gleiche gilt für Hypothesen über Trends (siehe z.B. Winer, 1971, 177-185; Hays, 1977, 687-694).

Eine Alternative zu den adjustierten multiplen t-Tests stellen multiple Vergleichstechniken dar, bei denen die kritischen Werte für die Einzeltests von vornherein so gewählt werden, daß die Wahrscheinlichkeit für mindestens ein fälschlich signifikantes Ergebnis in einer bestimmten „Familie von Tests“ einen festzulegenden Wert  $\alpha_G$  nicht überschreitet. Die bekanntesten dieser Techniken stammen von Scheffé (1959; vgl. Gabriel, 1964, 1969, 1978; Boik, 1979) und Tukey (1953; zit. n. Ryan, 1959; vgl. Gabriel, 1969; Williams, 1974, Keselman & Rogan, 1977, 1978; Keselman, Garnes & Rogan, 1979b). Beim Scheffé-Test wird der Rejektionsbereich so bestimmt, daß bei Testung aller möglichen Kontraste die Wahrscheinlichkeit mindestens eines Fehlers 1. Art höchstens gleich  $\alpha_G$  ist, bei den Tukey-Tests bezieht sich dieser Wert auf alle möglichen Kombinationen je zweier Mittelwerte.

Wir können auf diese und die anderen vorliegenden multiplen Vergleichsmethoden hier nicht näher eingehen. Leicht verständliche Darstellungen der gebräuchlichsten Verfahren geben Kirk (1968, S. 65-98), Dunnett (1970), Keppel (1973, Kap. 8) und Diehl (1979, Kap. 3). Des weiteren verweisen wir auf die Übersichten von Miller (1966, 1977), Garnes (1971 b, 1978a, b), O'Neill & Wetherill (1971), Lippman & Taylor (1972), Thomas (1973), Hopkins & Anderson (1973) sowie von Spjøtvoll (1974); zur Anwendung der Verfahren bei ungleichen Stichprobengrößen siehe u.a. Garnes (1971a), Smith (1971), Shaffer (1974a), Keselman, Toothaker & Shooter (1975), Keselman, Murray & Rogan (1976), Hochberg (1976) und Tamhane (1977). Vergleichende Untersuchungen zu einzelnen Verfahren auch hinsichtlich ihrer Robustheit gegenüber Verletzungen der Annahmen findet man etwa bei Ury (1971), Ury & Wiggins (1971, 1975), Ramseyer & Tcheng (1973), Keselman & Toothaker (1974), Howell & Garnes (1974), Garnes & Howell (1976), Keselman (1976), Kohr & Garnes (1977), Rogan, Keselman & Breen (1977), Keselman & Rogan (1977, 1978), Tamhane (1979); zur Teststärke dieser Verfahren siehe ferner Harter (1957), Steffens (1970), David, Lachenbruch & Brandis (1972), Einot & Gabriel (1975), Ramsey (1978).

Als Anhaltspunkte für die Wahl eines angemessenen Analyseverfahrens seien die folgenden Aspekte hervorgehoben:

- (1) Multiple t-Tests und der Scheffé-Test können zur Prüfung jedes *beliebigen* Mittelwertskontrasts herangezogen werden. Dabei sind multiple t-Tests effizienter, sofern die Zahl der zu prüfenden Vergleiche nicht wesentlich höher ist als die Zahl der Mittelwerte (vgl. u.a. Miller, 1966, 69; Garnes, 1978a).
- (2) Die Verfahren von Tukey sind für *paarweise* Mittelwertvergleiche effizienter als multiple t-Tests, es sei denn, es werden nur vglw. wenige der insgesamt möglichen Vergleiche durchgeführt (Dunn, 1961; Miller, 1966, 69; Garnes, 1978a).
- (3) Multiple t-Tests ermöglichen als einziges der genannten Verfahren stets auch die (teststärkere) Prüfung gerichteter Hypothesen.
- (4) Multiple t-Tests sind einfacher durchzuführen als die anderen Verfahren; ihre Teststärke ist problemlos bestimmbar; sie sind robust gegenüber Ver-

letzungen der zu ihrer validen Anwendung notwendigen Voraussetzungen, und daneben liegen vergleichbar einfache alternative Verfahren (siehe Kohr & Garnes, 1977; Garnes, 1978b) sowie relativ effiziente nicht-parametrische Analoga vor, deren Einsatz etwa bei mangelndem Skalenniveau indiziert ist.

Besonders in Anbetracht dessen, daß in hypothesen- oder theoriengeleiteten Untersuchungen nur relativ wenige Vergleiche von Interesse sein dürften (relativ zur Zahl möglicher (paarweiser) Kontraste), sollten in der Regel multiple t-Tests mit adjustierten  $\alpha$ - und  $\beta$ -Wahrscheinlichkeiten durchgeführt werden. Dies gilt unabhängig davon, ob die Mittelwertsvergleiche vor Datenerhebung geplant sind (s.o. Fall 2) oder nicht (Fall 1) und auch unabhängig davon, ob die Vergleiche statistisch unabhängig (orthogonal) sind oder nicht (siehe dazu Ryan, 1959; Davis, 1969; Ury & Wiggins, 1971, 1974; Myers, 1972, 362; Rodger, 1973). Erforderlich ist lediglich, daß vor der Durchführung der Mittelwertsvergleiche deren Anzahl festgelegt wird, um  $\alpha$  und  $\beta$  für jeden Einzeltest entsprechend adjustieren zu können.

### 8.3.2 Monotone Trendhypothesen

Wir betrachten nun den in der Psychologie sehr häufigen Fall, daß sich aus der wissenschaftlichen Hypothese eine statistische Hypothese über die Ordnung von Mittelwerten ableiten läßt (vgl. Abschn. 8.1.1):  $H_1: \mu_1 < \mu_2 < \dots < \mu_K$ . Diese Hypothese kann falsifiziert werden, wenn die Nullhypothese einer einfachen Varianzanalyse angenommen wird. Die Varianzanalyse kann also als Overall-Test zur Prüfung der hier betrachteten Hypothese benutzt werden. Wird die Alternativhypothese der Varianzanalyse angenommen, müssen zur Prüfung von  $H_1$  aber noch  $K-1$  paarweise Mittelwertsvergleiche mit den Alternativhypothesen  $H_{1(1)}: \mu_1 < \mu_2$  bis  $H_{1(K-1)}: \mu_{K-1} < \mu_K$  durchgeführt werden.

Als Entscheidungsregel kann man vereinbaren, daß die Gesamt-Alternativhypothese  $H_1$  angenommen werden soll, wenn alle Einzel-Alternativhypothesen  $H_{1(k)}$  (mit  $k = 1, \dots, K-1$ ) angenommen worden sind, wobei selbstverständlich auch abgeschwächte Kriterien denkbar sind. Die Wahrscheinlichkeit  $\alpha$  für die einzelnen Tests kann dann durchaus auch größer als  $\alpha = 0,05$  angesetzt werden. Sollen bspw. fünf Paar-Vergleiche durchgeführt werden, könnte man  $\alpha$  auf 0,20 festlegen, denn dann ist die Wahrscheinlichkeit einer Annahme der Gesamt- $H_1$ , obwohl alle Einzel- $H_{1(k)}$  falsch sind, mit  $0,20^5 = 0,0003$  extrem gering. Aber auch für den Fall, daß nur eine der  $H_{1(k)}$  falsch sein sollte, scheint eine Irrtumswahrscheinlichkeit von 0,20 für diese der wissenschaftlichen Hypothese doch noch recht gut entsprechenden Situation tolerabel zu sein.

Dagegen sollte die Wahrscheinlichkeit eines Fehlers 2. Art für jeden Einzeltest sehr klein sein, da es sonst leicht zu mindestens einer fälschlichen Annahme

einer  $H_{0(k)}$  und damit zu einer fälschlichen Ablehnung der Gesamt- $H$ , kommt. Daraus folgt die Notwendigkeit, hier möglichst effiziente statistische Tests zu verwenden. Nach dem im Abschnitt 8.3.1 Gesagten sind zur Prüfung monotoner Trendhypothesen also multiple t-Tests (oder entsprechende nicht-parametrische Verfahren) geeigneter als spezielle multiple Vergleichstechniken wie die Methoden von Scheffé und Tukey.

Als Overall-Test können anstelle der einfachen Varianzanalyse Verfahren angewendet werden, die von einer gerichteten Alternativhypothese ausgehen, d.h. nur dann zu signifikanten Resultaten führen, wenn aufgrund des Stichprobenergebnisses mindestens ein Mittelwertsunterschied in der spezifizierten Richtung besteht.

Ein entsprechendes parametrisches Verfahren stammt von Bartholomew (1959a, b, 1961a, b; s.a. Barlow et al., 1972); nicht-parametrische Verfahren sind für unabhängige Stichproben der Si-Test von Jonckheere (1954a, b; vgl. zur Darstellung und Beurteilung etwa Pm-i, 1965; May & Konkin, 1970; Odeh, 1972; Lienert, 1973, 279f.; Nelson & Toothaker, 1975 und Berenson, 1978) und für Designs mit einer Block- oder Kontrollvariablen der L-Test von Page (1963; vgl. zur Darstellung etwa Lienert, 1973, 357-360).

Diese Verfahren sind bei der hier betrachteten Fragestellung in der Regel effizienter als der varianzanalytische F-Test bzw. die entsprechenden Rangvarianzanalysen (Bartholomew, 1961a, b; Boersma, deJonge & Steilwagen, 1964; Puri, 1965; Garnes, 1966; Nelson & Toothaker, 1975). Da beim Overall-Test  $\beta$  zur Vermeidung ungerechtfertigter Falsifikationen sehr klein gewählt werden muß, ist die Verwendung dieser Tests in der hier betrachteten Situation von Vorteil.

Eine aus der psychologischen Hypothese abgeleitete statistische Hypothese über die Rangordnung von Mittelwerten kann selbstverständlich auch durch die ausschließliche Anwendung eines dieser Trendtests oder aber auch mit einem (Rang-)Korrelationstest geprüft werden. Da aber der Test jedes einzelnen Unterschiedes zwischen „benachbarten“ Mittelwerten zu einer insgesamt strengeren Prüfung der  $W$  führt, ist diese verfahrensweise i.a. vorzuziehen (vgl. Hager & Westermann, im Druck, a).

## 8.4 Mangelnde Präzision

Beim Signifikanztest erfolgt die Entscheidung über die Gültigkeit statistischer Hypothesen, indem der empirische Wert der Zufallsvariablen „Teststatistik“ in Beziehung zur Prüfverteilung gesetzt wird, d.h. zur Verteilung dieser Zufallsvariablen bei Gültigkeit der Nullhypothese (vgl. Abschn. 7.2). Die Wurzel aus der Varianz dieser Verteilung wird „Standardfehler“ der Teststatistik genannt (s. insbesondere Guilford & Fruchter, 1973). Wir können jetzt allgemein definieren: *Die Präzision der Prüfung einer statistischen Hypothese (oder einfacher: Die Präzision eines Experiments) ist um so höher, je kleiner der Standardfehler der entsprechenden Teststatistik ist.*

Bei festen Größen für die Wahrscheinlichkeit eines Fehlers 1. Art und für den in absoluten Einheiten ausgedrückten experimentellen Effekt (z.B.  $\alpha = 0,05$ ,  $EE = \mu_1 - \mu_2 = 3$ ) besteht folgende Beziehung: *Je geringer der Standardfehler der Teststatistik ist, desto geringer ist die Wahrscheinlichkeit  $\beta$  für einen Fehler 2. Art.* Deshalb ist (unter sonst gleichen Bedingungen) die statistische Validität eines Experimentes um so höher, je größer seine Präzision ist.

Welche Faktoren bestimmen die Größe des Standardfehlers von Teststatistiken und damit die Präzision von Experimenten?

- (1) Der Standardfehler einer Teststatistik ist (unter sonst gleichen Bedingungen) stets um so geringer, je größer die Zahl  $N$  der erhobenen Werte der abhängigen Variablen ist. (Zu den Folgerungen aus dieser Beziehung im Hinblick auf die Wahl der Stichprobengröße siehe Abschnitt 10.1.)
- (2) Bei parametrischen Testverfahren ist der Standardfehler einer Teststatistik (wiederum unter sonst gleichen Bedingungen) um so geringer, je kleiner die Varianz  $\sigma_e^2$  der sog. „unsystematischen Fehler“ ist, d.h. der Abweichungen der einzelnen Werte auf der abhängigen Variablen (in der Population) von ihrem Mittelwert (vgl. z.B. Lindquist, 1953, S. 2-5; Bredenkamp, 1969a, S. 339 und Abschn. 8.2.3). Deshalb wird die Präzision eines Experiments erhöht, wenn die Varianz  $\sigma_e^2$  der abhängigen Variablen in den Behandlungspopulationen verringert wird. Im folgenden werden wir eine Reihe von Techniken besprechen, die auf diese Weise zur Präzisionserhöhung und damit auch zu einer Vergrößerung der statistischen Validität beitragen können.

#### 8.4.1 Parallelisierung als Kontrolltechnik (StatV)

Besteht die Vermutung, daß die Varianz der AV durch eine ganz bestimmte individuelle Variable  $C$  beträchtlich erhöht wird und lassen sich Informationen über die Ausprägungen aller Probanden auf dieser Variablen beschaffen, so kann die Präzision des Experiments erhöht werden, indem diese Variable  $C$  als Kontrollfaktor ins Design eingeführt wird. Man spricht dann von einer „Parallelisierung“ (matching) der Probanden hinsichtlich dieser Variablen. Die Vorgehensweise sei am Beispiel eines Designs mit Ausprägungen einer UV erläutert.

Soll etwa hinsichtlich der Variablen „Intelligenzquotient“ parallelisiert werden, sind zunächst alle Probanden gemäß ihres IQ in eine Rangordnung zu bringen. Dann werden jeweils  $K$  Probanden der Reihe nach zu einem sog. „Block“ zusammengefaßt; die Zahl der Probanden muß also ein ganzzahliges Vielfaches von  $K$  sein. Innerhalb der  $Q=N/K$  Blöcke wird jeder experimentellen Bedingung dann genau ein Proband zufällig zugeordnet; daher rührt die Bezeichnung „*Plan der Zufallsblöcke*“ für diesen Versuchsplan 10.

(Vpl. 10)

		R-UV B				
		B <sub>1</sub>	...	B <sub>m</sub>	...	B <sub>K</sub>
(Kontroll- (Block-) variable) B-UV C	C <sub>1</sub>					
	⋮					
	C <sub>r</sub>			Y <sub>rm</sub>		
	⋮					
	C <sub>Q</sub>					

Zur Auswertung derartiger Pläne orientiere man sich in der in den Abschnitten 7.5.3.2, 7.5.3.3 und 8.4.6 angegebenen Literatur.

Eine Parallelisierung führt allerdings nur dann zu einer Erhöhung der Präzision des Experiments (d.h. zu einer Verringerung der Varianz der Prüfverteilung), wenn die Kontrollvariable eine zumindest mäßige statistische Assoziation mit der abhängigen Variablen aufweist - siehe dazu etwa Feldt (1958) sowie zur eingehenderen Begründung z.B. Myers (1972, 153-156).

Empfehlenswert ist es, mehrere benachbarte Blöcke zusammenzufassen, so daß in jeder Zelle des Versuchsplans 10 mehrere Werte der AV Y stehen. Dann kann nämlich auch die Interaktion zwischen der unabhängigen Variablen B und dem Kontrollfaktor C geprüft werden (vgl. Bredenkamp, 1969a). Feldt (1958) hat die Anzahl von Modalitäten des Kontrollfaktors bestimmt, die zu einer möglichst großen Präzisionserhöhung führen. Diese optimale Anzahl ist um so größer, je höher die Korrelation zwischen der Kontroll- und der abhängigen Variable, je größer die Gesamtzahl N der Beobachtungen und je kleiner die Zahl K der experimentellen Bedingungen ist.

Eine Parallelisierung mit mehreren experimentellen Einheiten pro Zelle des Designs ist auch durchführbar, wenn hinsichtlich der Kontrollvariablen C die Probanden nur grob in Klassen eingeteilt werden können (z.B. hohe, mittlere, niedrige Intelligenz) oder wenn C gar nur eine Klasseneinteilung ohne Definition einer Rangordnung darstellt (z.B. männlich/weiblich). Eine Parallelisierung ist nicht auf Experimente mit einer einzigen experimentellen unabhängigen Variablen beschränkt, vielmehr können die Vpn innerhalb eines Blocks natürlich auch zufällig auf die möglichen Kombinationen der Ausprägungen mehrerer Behandlungsvariablen aufgeteilt werden. Generell ist bei der Inter-



pretation und Analyse von Versuchsplänen mit Parallelisierung aber zu beachten, daß die Block- oder Kontrollvariable einer organismischen Variablen (im weitesten Sinn) entspricht, deren Ausprägung für jeden Probanden „vorgegeben“ ist. Wegen der fehlenden Zufallsordnung ist es nicht möglich, den statistischen Zusammenhang zwischen solchen Kontrollvariablen und der abhängigen Variablen kausal zu interpretieren (vgl. Abschn. 3.3).

#### 8.4.2 Kovarianzanalyse als Kontrolltechnik (StatV)

Bei der Parallelisierung wird die Fehlervarianz dadurch verringert, daß die mutmaßlich mit der AV verbundene Kontrollvariable als zusätzlicher Faktor in ein varianzanalytisches Design eingeführt wird. Bei der Kovarianzanalyse wird das gleiche Ziel zu erreichen versucht, indem man die Kontrollvariable C als unabhängige Variable im Sinne der Regressionsrechnung betrachtet. Die Kovarianzanalyse ist praktisch nichts anderes als eine Varianzanalyse, bei der aus den individuellen Unterschieden der Vpn auf der AV derjenige Teil herauspartialisiert wird, der mit Hilfe der linearen Regression auf die unterschiedlichen Werte auf der Kontrollvariablen zurückgeführt werden kann. Einzelheiten hierzu finden sich in der im Abschnitt 7.5.3.3 genannten Literatur sowie in den gesonderten Darstellungen der Kovarianzanalyse von Federer (1955, 483-522), Cochran (1957), Röhr (1975), Wildt & Ahtola (1976), Burnett & Barr (1977), Huitema (1980).

Gegenüber einer Varianzanalyse müssen bei Anwendung der Kovarianzanalyse mehrere zusätzliche *Voraussetzungen* erfüllt sein: Die Kontrollvariable muß quantitativ erfaßt sein, und innerhalb aller Behandlungsgruppen müssen die Regressionskoeffizienten sowie die Varianzen der Restvariablen  $Y^*$  gleich sein. (Zur Prüfung der statistischen Voraussetzungen siehe u.a. Kirk, 1968, 469-471 und Winer, 1971, 772-775; zu den Konsequenzen der Verletzung der Annahmen siehe Elashoff, 1969; Glass, Peckham & Sanders, 1972; Hamilton, 1976, 1977; Hollingsworth, 1980.)

Die statistischen Voraussetzungen sind stets dann nicht erfüllt, wenn man die Werte auf der Kontrollvariablen nach der experimentellen Behandlung ermittelt und wenn die Behandlungen sich unterschiedlich auf diese Werte auswirken (vgl. Evans & Anastasio, 1968; Sprott, 1970). Aus diesem Grund können mit Hilfe der Kovarianzanalyse auf keinen Fall Störfaktoren der internen Validität ausgeschaltet werden, d.h. ihre Anwendung kann einzig und allein der Erhöhung der Präzision dienen. (Zur Fehlanwendung der Kovarianzanalyse siehe weiter Lord, 1967; Harris, Bisbee & Evans, 1971; Overall & Woodward, 1977a, b.)

Da selbst bei Erfüllung der genannten Voraussetzungen die Kovarianzanalyse nur unter selten gegebenen Umständen zu einer wesentlich stärkeren Präzi-

sionserhöhung führt als eine Parallelisierung nach der entsprechenden Variablen (Cochran, 1957; Cox, 1957; Feldt, 1958), ist der Parallelisierung in der Regel der Vorzug zu geben. Mitunter kann aber deshalb nur eine Kovarianzanalyse durchgeführt werden, weil die Zeitspanne zwischen Erhebung der Kontrollvariablen und experimenteller Behandlung zu kurz ist, um parallele Gruppen zu bilden.

#### 8.4.3 Homogenisierung als Kontrolltechnik (StatV)

Wenn die Varianz  $\sigma_c^2$  zum Teil durch Unterschiede zwischen den Untersuchungseinheiten bedingt ist, kann sie verringert werden, indem man nur Vpn verwendet, die hinsichtlich einer oder mehrerer mit der AV zusammenhängenden Merkmalen homogen sind. Beispiel: Durch Prüfung unserer Hypothese  $WH_u$  ausschließlich an Personen mit (nahezu) gleichen Ausprägungen auf den Variablen „politisches Interesse“, „Alter“, „Art der Erfahrung mit Gastarbeitern“ und „Autoritätsgläubigkeit“ wird die Populationsvarianz in den Behandlungsgruppen wahrscheinlich verringert und die Präzision erhöht. Allerdings wird durch Beschränkung auf eine homogene Teilpopulation die Populationsvalidität stets vermindert. Die beste Lösung dieses Problems stellt die *systematische Variation* der vordem konstantgehaltenen Variablen dar. So könnte z.B. eine Untersuchung an mehreren, möglichst verschiedenartigen, in sich aber homogenen Probandengruppen durchgeführt werden, wodurch sowohl eine hohe Präzision als auch eine gute Populationsvalidität zu erreichen ist.

#### 8.4.4 Konstanthaltung und Elimination als Kontrolltechniken (StatV)

Im Abschnitt 3.3 haben wir Konstanthaltung und Elimination als Techniken zur Sicherung der internen Validität vorgestellt. Wir werden jetzt sehen, daß sie auch der Erhöhung der Präzision dienen können. Die Fehlervarianz  $\sigma_c^2$  entsteht auch durch Unterschiede hinsichtlich der genauen Bedingungen, unter denen die Werte der AV für jede experimentelle Einheit zustandekommen. Je stärker sich Bedingungen wie Darbietungszeit, Versuchsleiterverhalten, Lärmpegel, Gruppengröße, Tageszeit usw. für die einzelnen Vpn unterscheiden, desto größer wird  $\sigma_c^2$ . Die Präzision eines Experiments läßt sich also dadurch erhöhen, daß man die Unterschiede in solchen situationalen Variablen minimiert oder daß man diese Variablen idealerweise sogar konstant hält (z.B. gleiche Darbietungszeiten) bzw. ganz eliminiert (z.B. Ausschaltung einer Lärmquelle).

Diese Konstanthaltung ist im Laboratorium leichter möglich als in „natürlichen“ Situationen. Deshalb sind Experimente in der Regel präziser als Feldexperimente; anders ausgedrückt: Zur Entdeckung eines in absoluten Einhei-

ten gemessenen experimentellen Effektes sind im Labor weniger Beobachtungen notwendig als „im Felde“. Auf der anderen Seite schränkt jede Konstanthaltung von (situationalen) Bedingungen die Situationsvalidität ein (vgl. Abschnitt 4.2). Im Interesse einer möglichst strengen Prüfung der betrachteten Hypothese ist es daher oft ratsamer, diese Bedingungen zufällig variieren zu lassen und die Zahl der Beobachtungen dementsprechend zu erhöhen.

#### 8.4.5 Eingenistete Faktoren als Kontrolltechnik (StatV)

Eine weitere Möglichkeit zur Erhöhung der Präzision liegt in der Verwendung von hierarchischen Versuchsplänen, d.h. von Designs, bei denen die abhängige Variable nicht unter allen (theoretisch) möglichen Kombinationen von Modalitäten der unabhängigen Variablen beobachtet wird. Ein einfaches Beispiel möge dies verdeutlichen (vgl. Vpl. 11). Es sollen Hypothesen über die Wirksamkeit (AV Y) zweier Therapiemethoden  $A_1, A_2$  in Kombination mit drei Arten flankierender Maßnahmen  $B_1, B_2, B_3$  geprüft werden. Nehmen wir an, es gäbe drei Therapeuten  $C_1, C_2, C_3$ , die die Methode  $A_1$  durchführen können, und drei *andere* Therapeuten  $C_4, C_5, C_6$ , die Methode  $A_2$  beherrschen. Da die Modalitäten des Faktors C („Therapeuten“) nicht mit allen Modalitäten des Faktors A kombiniert werden können, bezeichnet man C als „in A eingenistet“ oder „geschachtelt“. Zur Prüfung der interessierenden Hypothesen wäre nur ein zweifaktorielles Design (vgl. Vpl. 6 in Abschn. 3.3.2) notwendig. Die Berücksichtigung des Faktors C hat aber (auch wenn er „eingenistet“ werden muß) den Vorteil, daß dadurch die Fehlervarianz verringert wird, die Präzision der Untersuchung also erhöht werden kann.

(Vpl. 11)

		R - U V B		
		$B_1$	$B_2$	$B_3$
R - U V A	$A_1$ N-UV C	$C_1$		
		$C_2$		
		$C_3$		
	$A_2$ N-UV C	$C_4$	$Y_{i422}$	
		$C_5$		
		$C_6$		

„N“ bedeutet „eingenistet“

Große Vorsicht ist bei der Verwendung hierarchischer Designs aber geboten, falls man auch über den eingestützten Faktor Hypothesen prüfen will. Dies ist nur unter der Annahme möglich, daß bestimmte Interaktionen (in Vpl. 11  $A \times C$  und  $A \times B \times C$ ) gleich Null sind (siehe dazu und zur Auswertung Winer, 1971, 359-366 und Bortz, 1979, 493-505).

#### 8.4.6 Wiederholte Messungen als Kontrolltechnik (StatV)

Wir wollen unter der gewählten Überschrift, die den häufigsten Verwendungszweck von Meßwiederholungen enthält, einige wesentliche Aspekte derartiger Pläne im Zusammenhang behandeln und insbesondere auf Auswertungsprobleme eingehen.

Meßwiederholungs- oder „Within-Subjects“-Designs werden in der einschlägigen Literatur unter verschiedenen Schwerpunktsetzungen und Überschriften behandelt, etwa als „Cross-over“- , „Change-over“- , „(multiple) Zeitreihen“- oder auch „Vor-Nachtest-Pläne“ (vgl. Hedayat & Afsarineyad, 1975) sowie „Split-Plot“-Pläne (siehe Kirk, 1968, Kap. 8; Winer, 1971, Kap. 7) oder „gemischte Pläne“ (Lindquist, 1953, Kap. 18).

Gemeinsam ist all diesen Versuchsanordnungen, daß von einer Beobachtungseinheit (Vp) mehr als ein Beobachtungsdatum erhoben wird. Hierbei sind in Abhängigkeit von der WH zwei Situationen zu unterscheiden:

(1) Jede Vp „liefert“ mehrere Scores auf mehreren  $AV_n$   $Y_q$  („multiple criteria“); die Vpn müssen nicht notwendigerweise mehreren Treatments unterzogen werden. Derartige Versuchspläne sind ihrer Natur nach multivariat und dienen sehr oft der intraexperimentellen Replikation eines theoretischen Konzeptes (vgl. Abschnitt 2.2). Beziehen sich die wissenschaftlichen Hypothesen auf die Relation zwischen der (den)  $UV(n)$  und den  $AV_n$  und ferner auf die Wechselwirkungen zwischen den letzteren, sollte eine multivariate Auswertung vorgenommen werden, da nur durch diese die Interaktionen zwischen den  $AV_n$  erfaßt werden können - ein instruktives Beispiel findet sich etwa bei Gabriel & Glavin (1978).

Interessieren dagegen nur die Wirkungen der  $UV_n$  auf die einzelnen  $AV_n$  und/oder interagieren diese nicht oder nicht nennenswert, sollte die Auswertung i.a. über mehrere univariate Tests erfolgen, denen zur Kontrolle der Fehlerwahrscheinlichkeiten ein globaler multivariater Test vorangestellt werden kann. Bei dieser Vorgehensweise werden jedoch auch andere statistische Hypothesen geprüft als mit den multivariaten Verfahren - siehe Gabriel & Hopkins (1974) und Bredenkamp (1980, 87).

(2) Jede Vp wird nur hinsichtlich einer AV beobachtet und gemessen, aber sukzessiv unter mehreren Modalitäten der UV, so daß pro Vp ebenfalls mehr

als ein Wert resultiert (vgl. Abschnitt 3.1). Außer in den Fällen, in denen sich derartige Pläne aus der zu prüfenden wissenschaftlichen Hypothese ergeben, werden sie häufig angewandt, um die Anzahl der bei interindividueller Bedingungsvariation benötigten Vpn zu reduzieren und gleichzeitig die Präzision zu erhöhen.

Letzteres geschieht, indem durch den Versuchsplan (und ein entsprechend adaptiertes statistisches Ausgangsmodell) die Trennung der Varianz zu Lasten der individuellen Merkmale von der zu Lasten der Fehleranteile erfolgt, die ja beide bei interindividueller Bedingungsvariation zur Fehlervarianz  $\sigma_e^2$  zusammengefaßt werden. Es resultiert eine für den statistischen Test der Treatmenteffekte um diesen individuellen Anteil reduzierte Prüfvarianz, sofern die übrigen Bedingungen unverändert bleiben - siehe dazu unten.

Wiederholte Messung als Technik zur Erhöhung der Präzision der Untersuchung kann als Spezialfall einer Parallelisierung betrachtet werden, bei der die Modalitäten des Kontrollfaktors den verschiedenen experimentellen Einheiten entsprechen (Vpl. 12).

(Vpl. 12)		W-UV B				
		B <sub>1</sub>	...	B <sub>m</sub>	...	B <sub>K</sub>
Vpn-UV V (Untersuchungs- einheiten- (Versuchs- personen-) Variable)	V <sub>1</sub>					
	⋮					
	V <sub>s</sub>			V <sub>sm</sub>		
	⋮					
	V <sub>n</sub>					

„W“ bedeutet UV mit wiederholter Messung

Eine Erweiterung des dargestellten Versuchsplanes ergibt sich, wenn man Meßwiederholungen über mehr als eine UV durchführt. In diesem Fall wird jede experimentelle Einheit unter allen möglichen Bedingungskombinationen beobachtet.

Von „gemischten Plänen“ spricht man dann, wenn in einem Experiment zugleich die inter- und die intraexperimentelle Bedingungsvariation realisiert wird. Im einfachsten Fall werden etwa jeder der J Modalitäten der einen experimentellen UV, etwa A, n Untersuchungseinheiten zugeordnet, und jede die-

ser  $J \cdot n$  Einheiten wird unter allen  $K$  Modalitäten der anderen UV, etwa  $B$ , beobachtet (Vpl. 13).

Im Vergleich mit vollständig randomisierten Plänen liegt der Vorteil von Meßwiederholungsplänen darin begründet, daß - wie bereits erwähnt - *unter sonst gleichen Bedingungen* eine höhere Präzision bei einer geringeren Anzahl von Vpn erzielt werden kann.

Allerdings ist hierbei zu berücksichtigen, daß sich die Vpn bei Mehrfachmessungen in einer anderen Situation befinden, sich anders verhalten als solche Vpn, die einem Treatment nur einmal unterzogen und beobachtet werden - vgl. Greenwald (1976, 212). Dies führt u.a. zu dem meist unvermeidbaren Auftreten der in den Abschnitten 3.2 und 3.4 aufgeführten unerwünschten Sequenzeffekte wie Übertragung, Übung, Ermüdung und Sensibilisierung, durch die die interne Validität bei Mehrfachmessungen an den gleichen Vpn schwieriger zu gewährleisten ist als bei interindividueller Bedingungsvariation (vgl. dazu auch Winer, 1971, 516-518).

(Vpl. 13)

			W-UV B				
			$B_1$	...	$B_m$	...	$B_K$
	Vpn-UV $V_{(A_1)}$ $A_1$	$V_{(A_1)1}$					
		$\vdots$					
		$V_{(A_1)s}$					
		$\vdots$					
		$V_{(A_1)n}$					
R-UV A	Vpn-UV $V_{(A_1)}$ $A_1$	$\vdots$					
		$V_{(A_1)1}$					
		$\vdots$					
		$V_{(A_1)s'}$			$Y_{s'lm}$		
		$\vdots$					
	Vpn-UV $V_{(A_j)}$ $A_j$	$V_{(A_j)n'}$					
		$\vdots$					
		$V_{(A_j)1}$					
		$\vdots$					
		$V_{(A_j)s''}$					
	Vpn-UV $V_{(A_j)}$ $A_j$	$\vdots$					
		$V_{(A_j)n''}$					
		$\vdots$					
		$V_{(A_j)s''}$					
		$\vdots$					

Erlebacher (1977, 1978) hat in diesem Zusammenhang einen Versuchsplan und eine Auswertetechnik vorgeschlagen, mit denen festgestellt werden kann, inwieweit die Art der Bedingungsvariation (inter- vs. intraindividuell) die Beziehung zwischen UV und AV beeinflusst.

Insgesamt ist aus den aufgeführten Gründen von der routinemäßigen Erhebung von Meßwiederholungen mit den *ausschließlichen Zielen* der Präzisionssteigerung und/oder Versuchspersonenökonomie abzuraten.

Kann jedoch die wissenschaftliche Hypothese strenger mittels eines Meßwiederholungsplanes geprüft werden, sollte dieser auch realisiert werden. Je nach Art der zu prüfenden Hypothese(n) und den Voraussetzungen für bestimmte statistische Tests ist dann zwischen unterschiedlichen Analysestrategien zu wählen.

#### 8.4.6.1 Analyse von Zeitreihen und Veränderungsmessungen

(1) Die WH bezieht sich auf Unterschiede oder Veränderungen in Abhängigkeit vom Zeitfaktor, dessen Einfluß daher bei der Analyse explizit zu berücksichtigen ist.

In diesem Fall liegt ein „Zeitreihenplan“ vor, dessen Auswertung in der Regel mittels der speziellen, unter dem Namen „Zeitreihenanalysen“ zusammengefaßten Verfahren vorgenommen werden sollte, die auch bei multifaktoriellen und multivariaten Plänen zur Anwendung kommen können. Kurze einführende Darstellungen in diese Techniken findet man etwa bei Rasch, Enderlein & Herrendörfer (1973), Dierkes (1977), Petermann (1978), Cook & Campbell (1979) und bei Revenstorf (1979); zur ausführlicheren, teilweise aber auch wesentlich schwierigeren Darstellung informiere man sich u.a. bei Wetzell (1970), Kendall (1973, 1976), Brillinger (1975), Chatfield (1975), Glass, Willson & Gottman (1975) sowie bei Box & Jenkins (1976); nicht-parametrische Zeitreihenanalysen stellt Lienert (1978) dar.

(2) Die WH beziehe sich auf Veränderungen im Sinne von Zuwachsraten etc., wobei der Zeitfaktor nicht explizit berücksichtigt wird.

Analysiert man die Daten als Differenz- oder Zuwachsrate („gain scores“, „change scores“), sind zahlreiche Probleme zu beachten, über die im einzelnen die Reader von Harris (1963) und de Gruijter & van der Kamp (1975, Teil 11) sowie die Arbeiten von Cronbach & Furby (1970), Linn & Slinde (1977), Rennert (1977), Petermann (1978) sowie Petermann & Hehl (1979) informieren. Diesen Problemen kann man teilweise ausweichen, indem man eine re-

gressionsanalytische Auswertung vornimmt, wie sie etwa Cohen & Cohen (1975, 377-393) skizziert haben.

#### 8.4.6.2 Univariate und multivariate Analysen

Die WH beziehe sich auf Unterschiede im Verhalten der gleichen Vpn unter verschiedenen experimentellen Bedingungen, wobei oft der Einfluß des Zeitfaktors auf die AV etwa durch randomisierte Reihenfolgen der Behandlungen ausgeglichen oder eliminiert werden kann - siehe Abschnitt 3.4.

In Abhängigkeit von den Voraussetzungen, die im konkreten Fall erfüllt sind, kann man sich grundsätzlich zwischen den folgenden Strategien der Auswertung entscheiden:

- (1) Man berechnet exakte univariate F-Tests, die auf vglw. restriktiven Voraussetzungen beruhen - siehe Abschnitt 8.4.6.2.1.
- (2) Man berechnet approximative univariate F-Tests, die bzgl. der Verletzung einer wesentlichen Voraussetzung adjustiert sind - siehe Abschnitt 8.4.6.2.1.
- (3) Man führt multivariate Varianz- bzw. Regressionsanalysen durch, die auf weniger restriktiven Bedingungen beruhen - vgl. Abschnitt 8.4.6.2.2.
- (4) Man nimmt eine nicht-parametrische Auswertung der Daten vor, die an relativ schwächere Voraussetzungen gebunden ist - vgl. Abschnitt 8.4.6.2.3.

Wir gehen im folgenden kurz auf die verschiedenen Auswertungstechniken ein.

##### 8.4.6.2.1 Exakte und approximative univariate Tests

Die im Abschnitt 8.2 angesprochenen Voraussetzungen zur Durchführung von Varianz- bzw. regressionsanalytischen F-Tests gelten uneingeschränkt auch für Versuchspläne mit wiederholten Messungen. Diese sind vornehmlich dadurch gekennzeichnet, daß die Beobachtungen unter den verschiedenen Treatments nicht unabhängig voneinander sind, weil sie jeweils an den gleichen Vpn erhoben wurden (vgl. Versuchsplan 12 in Abschnitt 8.4.6).

Zur Prüfung von Hypothesen über den wiederholten Faktor B mit K Modalitäten (S.U.) wird vorausgesetzt, daß die  $n \cdot K$  Beobachtungen einer K-dimensionalen Normalverteilung mit dem Mittelwertsvektor  $\mu$  und der (Varianz-)Kovarianz-Matrix  $\Sigma$  entstammen. Die Populations-Kovarianz-Matrix  $\Sigma$  über alle Paare von (potentiellen) Beobachtungswerten unter Faktor B hat folgende allgemeine Form (vgl. etwa Kirk, 1968; Tatsuoka, 1971; Moosbrugger, 1978):



$$(8.10) \quad \Sigma = \begin{bmatrix} B_1 & B_2 & \dots & B_m & \dots & B_K \\ \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1m} & \dots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2m} & \dots & \sigma_{2K} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_m^2 & \dots & \sigma_{mK} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_{K1} & \sigma_{K2} & \dots & \sigma_{Km} & \dots & \sigma_K^2 \end{bmatrix} \begin{matrix} B_1 \\ B_2 \\ \dots \\ B_m \\ \dots \\ B_K \end{matrix}$$

Hypothesen über den Mittelwertsvektor  $\mu$  werden i.a. mittels des folgenden univariaten F-Tests geprüft (vgl. Winer, 1971, 281):

$$(8.11) \quad F = \frac{s_{\text{Treatment B}}^2}{s_{\text{p}n \times \text{Treatment B}}^2}; \quad df = (K - 1); \quad (n - 1)(K - 1)$$

Dieser Test ist bspw. dann valide, d.h. F ist unter Gültigkeit der Nullhypothese zentral F-verteilt, wenn die Kovarianz-Matrix folgende spezifische Struktur aufweist:

$$(8.12) \quad \Sigma = \begin{bmatrix} \sigma_e^2 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sigma_e^2 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_e^2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & \sigma_e^2 \end{bmatrix} = \sigma_e^2 \cdot I_K$$

$I_K$  bezeichnet die  $K \times K$  Einheitsmatrix; ferner ist  $\sigma_k^2 = \sigma_e^2$  und  $\sigma_{mm'} = 0$  für alle  $m \neq m'$ . Dies bedeutet, daß alle Beobachtungen unabhängig voneinander und daß alle Varianzen homogen sind (vgl. Abschnitt 8.4.2).

Die Werte des F-Bruches sind allerdings auch dann F-verteilt, wenn die Varianzen und die Kovarianzen homogen sind, wenn also gilt:

$$(8.13) \quad \Sigma = \begin{bmatrix} \sigma_e^2 & R \cdot \sigma_e^2 & \dots & R \cdot \sigma_e^2 & \dots & R \cdot \sigma_e^2 \\ R \cdot \sigma_e^2 & \sigma_e^2 & \dots & R \cdot \sigma_e^2 & \dots & R \cdot \sigma_e^2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ R \cdot \sigma_e^2 & R \cdot \sigma_e^2 & \dots & R \cdot \sigma_e^2 & \dots & \sigma_e^2 \end{bmatrix}$$

R bezeichnet die Populationskorrelation zwischen Beobachtungen an zwei beliebigen Meßzeitpunkten.

Weisen Matrizen die vorstehende Form auf, spricht man von „Uniformität“ („compound symmetry“) (vgl. Winer, 1971, 277; Davidson, 1972). Uniformität der Kovarianz-Matrix ist jedoch trotz anderslautender Behauptungen (e.g. Davidson, 1972; Poor, 1973; Keppel, 1973) keine notwendige, sondern lediglich eine hinreichende Bedingung für die Validität des F-Bruches (8.11).

Um die notwendigen und hinreichenden Bedingungen für die Validität des obigen F-Bruches zur Auswertung eines einfachen Meßwiederholungsplanes wie Versuchsplan 12 (s.o.) kurz skizzieren zu können, gehen wir davon aus, daß die  $n \cdot K$  Beobachtungen in Form einer  $n \times K$  Matrix angeordnet sind, deren  $n$  Zeilen die Beobachtungen jeder  $V_p$  auf den  $K$  Variablen (Modalitäten von  $B$ ) enthalten (siehe Versuchsplan 12 in Abschnitt 8.4.6). Die folgenden Definitionen beziehen sich auf diese Rahmenbedingungen und sind zum Verständnis der Voraussetzungen unerlässlich (vgl. Rouanet & Lepine, 1970).

Ein *Kontrast* zwischen den Treatmentbedingungen des wiederholten Faktors  $B$  bezeichnet einen Vektor  $c$ , dessen  $K$  Komponenten sich zu Null aufaddieren. Die Norm dieses Kontrastes ist definiert als  $(c' \cdot c)^{1/2}$ , und ein *standardisierter* Kontrast hat die Norm 1. Die  $K \times f$  Matrix  $C$  wird dann als *Kontrast-Matrix* bezeichnet, wenn ihre  $f$  Spaltenvektoren Kontraste darstellen. Sind die  $f$  Spaltenvektoren paarweise orthogonal, heißt  $C$  spaltenweise oder  $f$ -orthogonal, und sind zudem alle Spalten standardisiert, wird  $C$  *spaltenweise* oder *f-orthonormal* genannt.

Eine  $f$ -orthonormale  $K \times f$  Matrix kann daher bspw. die folgende Form aufweisen (vgl. Mendoza, Toothaker & Crain, 1976; Rogan, Keselman & Mendoza, 1979; Huynh & Mandeville, 1979):

$$(8.14) \quad C = \begin{bmatrix} 1/c_1 & 1/c_2 & \dots & 1/c_f \\ -1/c_1 & 1/c_2 & \dots & 1/c_f \\ 0 & -2/c_2 & \dots & 1/c_f \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -f/c_f \end{bmatrix}$$

mit  $c_g = g(g+1)^{1/2}$ ,  $g = 1, \dots, f$ .

Für eine Matrix der Form (8.14) gilt:

$$(8.15) \quad C'C = I_f, \text{ mit } I_f : f \times f \text{ Einheitsmatrix.}$$

Ein *Vergleich* zwischen den  $K$  Stufen des wiederholten Faktors  $B$  wird durch die Kontrast-Matrix  $C$  repräsentiert, und die in ihr enthaltenen  $f$  Kontrastvektoren können mit der Einschränkung beliebig gewählt werden, daß sie orthonormiert sein müssen (s. o.). Dabei bezeichnet  $f$  die Freiheitsgrade dieses Vergleichs (siehe Rouanet & Lepine, 1970; Rogan, Keselman & Mendoza, 1979, 275).

Rouanet & Lepine (1970, 152) unterscheiden nun u.a. die folgenden Arten von Vergleichen: Ein Vergleich mit  $df = 1$  wird *eindimensional* genannt; er kann durch einen einzigen Kontrast repräsentiert werden. Ein Vergleich mit  $df = f$  heißt *partial*, wenn  $f < K - 1$ , und *global* („Overall“), wenn  $f = K - 1$ .

Interessiert nun der *globale* Vergleich, ist die Kontrast-Matrix als  $K \times (K - 1)$  Matrix mit  $K - 1$  orthonormalen Spaltenvektoren zu konstruieren, und der entsprechende F-Test (8.11) ist genau dann valide, wenn für die Kovarianz-Matrix  $\Sigma$  gilt (vgl. Rouanet & Lepine, 1970; Huynh & Feldt, 1970):

$$(8.16) \quad C'\Sigma C = \sigma_e^2 I_{K-1}$$

M. a. W. : Können die  $K$  ursprünglichen Variablen, für die jeweils  $n$  wiederholte Messungen vorliegen, unter Verwendung einer geeigneten orthonormalen Kontrast-Matrix (s.o.) in  $K - 1$  orthonormale Variablen transformiert werden, die unabhängig voneinander und homogen variant sind, kann die statistische Signifikanz des wiederholten Faktors mittels F-Test valide beurteilt werden (Huynh & Mandeville, 1979, 965; Rogan, Keselman & Mendoza, 1979). In diesem Fall besitzt die Kovarianz-Matrix die Eigenschaft der „Zirkularität“ („circularity“, „sphericity“) *bzgl. des gewählten Vergleichs*, die die notwendige und hinreichende Bedingung für den o.a. F-Test darstellt (vgl. zum Beweis Rouanet & Lepine, 1970 sowie Huynh & Feldt, 1970).

Liegt der zweifaktorielle Versuchsplan 13 (s. o.) mit wiederholten Messungen auf einem Faktor (B) vor, ist mit jeder der  $J$  Stufen von A eine eigene Kovarianz-Matrix  $\Sigma_j$  verbunden. Die Tests für B und die Interaktion  $A \times B$  sind nur dann valide, wenn gilt, daß die  $J$  Kovarianz-Matrizen  $C'\Sigma_j C$  auf allen Stufen von A gleich sind und wenn die gemeinsame Matrix  $C'\Sigma C$  zirkulär ist („multisample circularity“; vgl. Huynh & Mandeville, 1979; Rogan, Keselman & Mendoza, 1979). In einem zweifaktoriellen Plan mit zwei wiederholten Faktoren B mit  $K$  Stufen und D mit  $M$  Stufen erfolgt die Testung von B gegen  $s_{V_{pn} \times B}^2$ , von D gegen  $s_{V_{pn} \times D}^2$  und von  $B \times D$  gegen  $s_{V_{PN} \times B \times D}^2$  (Eimer, 1978, 118ff.). Hier ist die Kovarianz-Matrix  $\Sigma$  vom Typ  $KM \times KM$ , und mit jedem Test ist eine eigene Kontrast-Matrix  $C$  verbunden. Diese ist für den Faktor B vom Typ  $KM \times (K - 1)$ , für den Faktor D vom Typ  $KM \times (M - 1)$  und für die Interaktion vom Typ  $KM \times (K - 1)(M - 1)$ . Mit jedem der drei durch seine zugehörige Kontrast-Matrix repräsentierten globalen Vergleiche ist eine eigene Zirkularitätsbedingung der Form (8.16) verknüpft; die drei Skalare  $\sigma_B^2$ ,  $\sigma_D^2$  und  $\sigma_{BD}^2$  müssen nicht gleich sein (siehe zu den Einzelheiten Mendoza, Toothaker & Crain, 1976; Huynh & Mandeville, 1978; vgl. auch Scheifley & Schmidt, 1978). In dieser Literatur finden sich auch Hinweise auf die Validitätsbedingungen in weiteren Versuchsplänen mit wiederholten Messungen auf einigen oder allen Faktoren.

Die Überprüfung der Zirkularität der Matrizen kann über die entsprechenden Tests nach Mauchly (1940) und Box (1950) erfolgen, die u.a. in Kirk (1968), Winer (1971), Morrison (1976) und Bortz (1979) dargestellt sind; eine Erweiterung des Tests von Mauchly (1940) hat Mendoza (1980) vorgeschlagen. Die genannten Verfahren sind jedoch nicht robust gegenüber Verletzungen der Annahme der multinormalen Verteilung der Fehlerterme und darüber hinaus

sehr teststark, weswegen i. a. die Hypothese der Zirkularität zurückgewiesen werden dürfte (vgl. Davidson, 1972; Rogan, Keselman & Mendoza, 1979). Die Anwendung dieser Verfahren zur Überprüfung der Validitätsbedingungen der F-Tests ist daher in der Regel nicht zu empfehlen, wie Keselman et al. (1980) ausführen und begründen.

Führt man ungeachtet der Struktur der Kovarianz-Matrizen univariate F-Tests durch,<sup>30)</sup> ist je nach Ausmaß der Abweichung von der oben spezifizierten Struktur der Kovarianz-Matrizen teilweise mit starkem Anwachsen der tatsächlichen Wahrscheinlichkeiten für Fehler 1. Art zu rechnen (Box, 1954 b; Wilson, K., 1975). Abweichungen von dieser Struktur scheinen darüber hinaus in der Empirie die Regel zu sein (Davidson, 1972; Keppel, 1973, 464; Wilson, R. S., 1975; Greenwald, 1976; Rogan, Keselman & Mendoza, 1979; vgl. in diesem Zusammenhang auch Wallenstein & Fleiss, 1979).

*Die verbreitete Auswertung von Meßwiederholungsplänen über konventionelle F-Tests ist daher insgesamt wenig empfehlenswert.*

Die Abweichung einer Kovarianz-Matrix von der Zirkularität hinsichtlich des globalen Vergleichs kann mittels des von Box (1954 b) abgeleiteten Index'  $\epsilon$  erfaßt werden, für den gilt:

$$(8.17) \quad \epsilon = \frac{(\text{tr } \mathbf{C}'\mathbf{\Sigma} \mathbf{C})^2}{(K-1) \text{tr}(\mathbf{C}'\mathbf{\Sigma} \mathbf{C})^2}, \text{ mit tr: Spur der genannten Matrix.}$$

(vgl. Rouanet & Lepine, 1970, 156; siehe ferner Geisser & Greenhouse, 1958; Greenhouse & Geisser, 1959; Keselman & Mendoza, 1979). Der Wert des Maßes  $\epsilon$  kann aus den empirischen Daten geschätzt werden - hierüber informieren Huynh & Feldt (1976) und Huynh (1978). Box (1954b) hat zeigen können, daß unter Verwendung von  $\epsilon$  die Freiheitsgrade des univariaten F-Tests so adjustiert werden können, daß ein approximativer, aber hinreichend valider Test selbst dann resultiert, wenn die Abweichungen von der spezifizierten Struktur der Matrizen sehr ausgeprägt sind (vgl. Collier et al., 1967; Davidson, 1972; Gaito, 1973; Mendoza, Toothaker & Nicewander, 1974; Wilson, K., 1975; Wilson, R. S., 1975; Rogan, Keselman & Mendoza, 1979).

*Diese approximativen Tests bieten sich daher zur Prüfung univariater Hypothesen an; siehe im einzelnen Huynh (1978).*

---

<sup>30)</sup> über die entsprechenden Tests informieren etwa Winer (1971, 261-273, 514-594), Myers (1972, 168-186), Keppel (1973, 401-408), Cohen & Cohen (1975, 403-412) und Eimer (1978); zur Kodierung der Versuchspersonen siehe Pedhazur (1977).

#### 8.4.6.2.2 Multivariate Tests

Die multivariaten Varianz- und Regressionsanalysen sowie verwandte Verfahren (vgl. Knapp, 1978) ermöglichen die Prüfung von Hypothesen über die unspezifische Wirkung von einer oder mehreren UVn auf mehrere AVn  $Y_q$  und die Wechselbeziehungen zwischen letzteren (vgl. Moosbrugger, 1978, 104). Ihre valide Anwendung ist u. a. mit der Annahme multinormal verteilter Modellresiduen verknüpft (siehe zu Einzelheiten etwa Moosbrugger, 1978, 121), wobei moderate Abweichungen tolerabel sind (e.g. Olson, 1974). Ferner müssen die Kovarianz-Matrizen für alle Bedingungen des nicht-wiederholten Faktors homogen sein, wobei allerdings keine spezifischen Annahmen bzgl. ihrer Struktur erforderlich sind (vgl. Olson, 1974; Bredenkamp, 1980, 84; Stevens, 1980; Rogan, Keselman & Mendoza, 1979). Insofern ist die valide Anwendung entsprechender multivariater Tests zur Auswertung von Meßwiederholungsplänen an insgesamt etwas schwächere Bedingungen geknüpft als die univariater Tests, wobei erstere allerdings auch bei gleichen Stichprobenumfängen anfälliger gegenüber Heterogenität der Kovarianz-Matrizen sind als die adjustierten univariaten Tests (Olson, 1974; Huynh, 1978).

Generell stellt sich bei multivariaten Analysen die Frage nach dem unter Robustheitsaspekten „optimalen“ Testkriterium; hierauf gehen auch die bekannteren Lehrbücher kaum ein (vgl. die Angaben im Abschnitt 7.5.3.3). Von den verschiedenen Simulationsstudien, mit denen eine Beantwortung dieser Frage versucht wurde, verdienen insbesondere die von Olson (1974, 1976, 1979) und Stevens (1979, 1980) Erwähnung; spezielle Robustheitsuntersuchungen haben in jüngerer Zeit Everitt (1979) sowie Hakstian, Roed & Lind (1979) durchgeführt. Aus Olsons und Stevens' Befunden geht hervor, daß i.a. die Verwendung des von Bartlett (1937, 1939) und Pillai (1955) entwickelten V-Kriteriums zur geringsten Anzahl falscher Entscheidungen führen dürfte, wenn man davon ausgeht, daß dem E die „wahre“ Populationssituation (e. g. die Art der Nicht-Zentralitäts-Struktur) unbekannt ist.

Zur Frage der Anschlußtests nach einem signifikanten globalen Test siehe insbesondere Morrison (1976) sowie Stevens (1972b), Spector (1977) und Ramsey (1980).

*Insgesamt stellen die multivariaten Ansätze eine Alternative zu den im übrigen nicht weniger rechenaufwendigen approximativen univariaten F-Tests dar. Die Entscheidung zugunsten einer der beiden Auswertetechniken sollte daher im Einzelfall vor allem davon abhängig gemacht werden, welche der jeweils geprüften statistischen Hypothesen am ehesten zu einer möglichst strengen Prüfung der wissenschaftlichen Hypothese führt.*

Bzgl. weiterer Einzelheiten siehe neben der bereits aufgeführten Literatur u.a. Hummel & Sligo (1971), Poor (1973) sowie Romaniuk, Herbert & Levin (1977).

#### 8.4.6.2.3 Nicht-parametrische Tests

Die adäquaten nicht-parametrischen Auswerteverfahren sind an wesentlich schwächere Voraussetzungen gebunden (vgl. Abschnitt 7.5.3). Allerdings stehen diese Techniken nur für vglw. einfache Meßwiederholungspläne zur Verfügung, so daß sie insgesamt kaum eine Alternative zu den parametrischen Tests darstellen dürften (vgl. McCall & Appelbaum, 1973), es sei denn, ein zu geringes Skalenniveau erzwingt ihren Einsatz. In diesem Fall können dann komplexe Pläne oft nicht analog ausgewertet werden. Bzgl. der Einzelheiten zu den Verfahren sei auf Lienert (1973, 345-369; 1978, 1045-1054), Siegel (1976, Kap. 7) und Marascuilo & McSweeney (1977, 354-388) verwiesen sowie ferner auf die im Abschnitt 7.5.3.1 genannte Literatur zur Auswertung nominaler Daten.

Abschließend sei nochmals hervorgehoben, daß die interne und möglicherweise auch die Variablen-Validität bei Meßwiederholungen in den meisten Fällen nicht hinreichend gesichert werden kann (vgl. auch die Abschnitte 3.2 und 3.4).

#### 8.4.7 *Zur Beziehung zwischen der Präzision und den anderen Aspekten der experimentellen Validität*

Wir sind in diesem Abschnitt 8.4 mehrfach darauf gestoßen, daß durch Techniken zur Erhöhung der Präzision andere Aspekte der experimentellen Validität beeinträchtigt werden können: Bei Konstanthaltung von Situationsvariablen wird die Situationsvalidität herabgesetzt, mit homogenen Versuchspersonengruppen hat man eine geringere Populationsvalidität, bei Meßwiederholung ist die interne Validität schwieriger zu gewährleisten, die Einführung von Kontroll- oder Kovariablen kann zu Erwartungseffekten führen, wenn diese Variablen in Vortests erhoben werden, usw.

In Lehrbüchern der Versuchsplanung wird häufig auch empfohlen, die Präzision zu erhöhen, indem ein komplexeres Design verwendet wird: Durch Einführung weiterer systematischer Faktoren kann den Daten ein jeweils komplexeres Modell angepaßt werden. Hierdurch wird es möglich, aus der ursprünglichen Fehlervarianz solche Anteile herauszulösen, die durch die zusätzlichen Faktoren „erklärt“ werden können - Ähnliches geschieht bei der Parallelisierung, Kovarianzanalyse und der Meßwiederholung. Gegen ein unreflektiertes Befolgen dieser Empfehlung sprechen jedoch mehrere Gründe:

1. Die Beziehung zwischen der Zahl der Faktoren und der Präzision des Experiments gilt „unter sonst gleichen Bedingungen“, d.h. u.a. bei gleichbleibender Stichprobengröße pro Zelle (Bedungskombination). Je mehr Faktoren untersucht werden, desto höher wird der Gesamtversuchspersonenbedarf, wenn man das  $n$  pro Zelle konstant halten will.
2. Je mehr Faktoren ein Versuchsplan umfaßt, desto stärker kumulieren bei dem gebräuchlichen Auswertungsvorgehen (Durchführung aller Tests) die Fehlerwahrscheinlichkeiten  $\alpha$  und  $\beta$ . Darüber hinaus erfolgt in derartigen Plänen die Prüfung der statistischen Hypothesen i.a. gegen nur eine oder aber nur wenige Prüfvarianzen. Dies hat zur Folge, daß die einzelnen F-Brüche nicht mehr unabhängig voneinander sind (Kimball, 1951; Hurlburt & Spiegel, 1976; Hays, 1977, 516f., 589f.). Beide Probleme lassen sich u.E. nur dann weitgehend vermeiden, wenn man die Anzahl der Faktoren ausschließlich nach dem Kriterium der strengen Prüfung der WH festlegt und sich bei der Auswertung auf die statistischen Hypothesen beschränkt, die im Hinblick auf die WH relevant sind (Hager & Westermann, im Druck, a, b).

Aus diesen Einwänden geht hervor, daß es keineswegs stets dem Ziel möglichst strenger Prüfungen von Kausalhypothesen dienen muß, wenn man durch die Anwendung von Kontrolltechniken das Experiment so präzise wie möglich macht. Daher lehnen wir auch die starre sog. „MaxMinKon-Strategie“ (vgl. Wormser, 1974, Kap. 5; Kerlinger, 1978, 450-459) ab, nach der ein Experiment *stets* u.a. so zu planen ist, daß durch die möglichst umfassende Kontrolle von unsystematischen Einflüssen die Fehlervarianz minimiert wird.

## 8.5 Falsche Analyse und Interpretation statistischer Interaktionen

Wir haben bereits erwähnt, daß sich Hypothesen über das Vorliegen statistischer Interaktionen zwischen (jeweils) zwei oder mehr Variablen unmittelbar aus der betrachteten wissenschaftlichen Hypothese oder Theorie ergeben können (siehe Abschnitt 8.1.1, Punkt (9)) und daß sie von Interesse sind, wenn untersucht wird, inwieweit ein vorhergesagter Unterschied zwischen Behandlungsbedingungen von der Ausprägung einer (oder mehrerer) (störender) Populations- oder Situationsvariablen abhängig ist (vgl. Abschnitt 4.3).

Über die Bedeutung, Prüfung und Interpretation statistischer Interaktionen herrscht nach unserem Eindruck einige Unklarheit, zumal man in nur ganz wenigen Lehrbüchern eine genügend differenzierte Darstellung findet (e. g. in Lindquist, 1953; Diehl, 1979; Henning & Muthig, 1979). Diese Unklarheit kann zu falschen Entscheidungen über die Gültigkeit wissenschaftlicher Hypothesen führen, weswegen wir im folgenden einige grundsätzliche Erläute-

rungen zu diesem Thema geben wollen. Hierbei beschränken wir uns auf die sog. „varianzanalytischen Interaktionen“ und gehen auf diejenigen, die als „Zusammenhang zwischen kategorialen Merkmalen“ definiert sind, nicht ein - nähere Einzelheiten hierzu entnehme man der im Abschnitt 7.5.3.1 angegebenen Literatur zur Auswertung nominaler Daten.

Ob man eine Interaktion erfassen kann, ist von der Art des statistischen Modells abhängig, das man den Daten anpassen will. Nicht immer sind Interaktionen empirisch sinnvoll interpretierbar oder aber von theoretischem Interesse. In einem solchen Fall kann der E ein Modell anpassen, das keine oder aber eine reduzierte Anzahl von Interaktionsparametern enthält. Dies hat oft den Vorteil, daß interessierende Hypothesen etwa mittels hierarchischer Pläne (vgl. Abschnitt 8.4.5) ökonomischer geprüft werden können; zur Frage der prüfbaren Hypothesen und der alternativen Modelle vgl. neben den Standardlehrbüchern der Versuchsplanung und -auswertung etwa Elston & Bush (1964), Marascuilo & Levin (1970, 1976), Levin & Marascuilo (1972, 1973), Garnes (1973, 1978c) sowie Betz & Gabriel (1978). Allerdings ist bei der Elimination von Interaktionsparametern bei der Modellanpassung stets die Bedeutung zu berücksichtigen, die Interaktionen für die Populations- und Situationsvalidität zukommen kann (vgl. dazu Abschnitt 4.3); wir greifen diesen Gedanken im folgenden auf.

Hat die AV Rangniveau, empfiehlt sich die nicht-parametrische Auswertung des entsprechenden Versuchsplanes über die Rangtests nach Kruskal & Wallis (1952) sowie Friedman (1937) (vgl. auch Lienert, 1973), mit denen allerdings keine Interaktionen erfaßt werden können. Um dies zu ermöglichen, hat Brendenkamp (1974, 1980, 73f.) bestimmte Modifikationen dieser Verfahren vorgeschlagen, auf die wir hier nicht eingehen - siehe zu Einzelheiten auch Gebert (1977), Krüger (1977) und Engelhardt (1979).

Die folgenden Ausführungen zur Interaktion bei intervallskalierter AV gelten im wesentlichen auch bei rangskalierter AV.

### *8.5.1 Das Konzept der statistischen Interaktion*

In einem Versuchsplan, mit dem die simultane Wirkung von mindestens zwei (quantitativen oder qualitativen) UVn auf die quantitative(n) AV(n) untersucht wird, können unabhängig vom Skalenniveau der AV stets statistische Interaktionen auftreten. Man bezeichnet häufig die Wechselwirkung zwischen (jeweils) zwei UVn als „Interaktion 1. Ordnung“, während die „Interaktionen höherer Ordnung“ durch das simultane Wirken von jeweils mehr als zwei UVn auf die AV(n) charakterisiert sind.



Die folgenden Darstellungen beschränken sich auf die Interaktionen 1. Ordnung, und wir gehen der Einfachheit halber von einem vollständig gekreuzten zweifaktoriellen varianzanalytischen Versuchsplan mit zwei UVn A und B mit J resp. K fixierten Modalitäten aus. Jeder der K-J Bedingungskombinationen AB wurden zufällig n Vpn zugewiesen. Wir nehmen an, daß die Voraussetzungen zur Durchführung einer Varianzanalyse via F-Tests gegeben sind.

In einem derartigen Experiment werden in jeder Zeile  $A_i$  K Mittelwerte  $M_{ik}$  und in jeder Spalte  $B_k$  J Mittelwerte  $M_{jk}$  auftreten, die als Stichprobenäquivalente der Modellparameter  $\mu_{jk}$  angesehen werden.

Man bezeichnet nun die Unterschiede zwischen allen Mittelwerten in einer Zeile  $A_i$  als den „einfachen Haupteffekt des Faktors B in der Modalität  $A_i$  des Faktors A“ oder kurz als „einfachen (Haupt-)Effekt (von) B in  $A_i$ “; entsprechend heißen die Unterschiede zwischen allen Mittelwerten der Spalte B, „einfacher (Haupt-)Effekt (von) A in B,“.

Der Durchschnitt der einfachen Haupteffekte eines Faktors X über alle Stufen des jeweils anderen Faktors, m.a. W. die Unterschiede zwischen den Rand- oder Marginalmittelwerten, nennt man den „Haupteffekt des Faktors X“, wobei „X“ entweder gleich A oder aber gleich B ist. Tab. 8.1 möge dies etwas veranschaulichen.

Tabelle 8.1: Einfache Effekte und Haupteffekte.

	Faktor B								Zeilen- mittelwerte
	$B_1$	$B_2$	...	...	$B_m$	.	.	.	$B_K$
Faktor A	$A_1$	einfacher Effekt B in $A_1$							
	$A_2$	einfacher Effekt B in $A_2$							
	...	...							
	$A_i$	einfacher Effekt B in $A_i$							
	...	...							
	$A_j$	einfacher Effekt B in $A_j$							
					einfacher Effekt A in $B_m$				einfacher Effekt A in $B_K$
							...		
									(über alle Spalten $B_k$ )
Spalten- mittel- werte	Haupteffekt B (über alle Zeilen $A_j$ )								Gesamt- mittel- wert

Unabhängig von der Gleichheit oder Ungleichheit der Randmittelwerte zu Lasten eines Haupteffektes können Unterschiede zwischen den Zellenmittelwerten  $\mu_{jk}$  bestehen, die *nicht* auf die Haupteffekte zurückführbar sind. Bestehen zwischen den Populationszellenmittelwerten  $\mu_{jk}$  derartige Unterschiede oder Differenzen, spricht man von „statistischer Interaktion“ oder „Wechselwirkung“. Lindquist (1953, 126) definiert einen Interaktionseffekt „als einen Unterschied zwischen einem einfachen Effekt und dem korrespondierenden Haupteffekt“. Eine Wechselwirkung liegt m.a. W. dann vor, wenn die einfachen Effekte A nicht auf allen Stufen des Faktors B gleich sind und/oder umgekehrt (vgl. dazu im einzelnen Lindquist, 1953; Lee, 1961; Lubin, 1961; Digman, 1966; Bracht & Glass, 1968; Bracht, 1970; Plomp, 1974; Bredenkamp, 1975).

Im Fall der hier vorausgesetzten Versuchsanordnung (S.O.) kann den Daten das folgende additive Modell angepaßt werden (vgl. dazu auch Abschnitt 8.2):

$$(8.18) \quad Y_{ijk} = \mu + (\mu_j - \mu) + (\mu_k - \mu) + (\mu_{jk} - \mu_j - \mu_k + \mu) + e_{ijk} \\ = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + e_{ijk}$$

Hierbei gilt:

$$(8.19) \quad \alpha_j = (\mu_j - \mu): \text{Haupteffekt A}$$

$$(8.20) \quad \beta_k = (\mu_k - \mu): \text{Haupteffekt B}$$

$$(8.21) \quad (\alpha\beta)_{jk} = (\mu_{jk} - \mu) - (\mu_j - \mu) - (\mu_k - \mu) = \\ = (\mu_{jk} - \mu_j - \mu_k + \mu): \text{Interaktionseffekt AB}$$

Man ersieht aus Gleichung (8.21), daß die Interaktion AB denjenigen Variationsanteil darstellt, der in den einzelnen Zellen des Designs verbleibt, nachdem von dem Abweichungsausdruck  $(\mu_{jk} - \mu)$  die Variation zu Lasten der beiden Haupteffekte A und B subtrahiert worden ist.

Der entsprechende varianzanalytische F-Test prüft die  $H_0$ , daß diese „Reste“, quadriert und aufsummiert, in der Population gleich Null sind, daß also keine Interaktion vorliegt. Fiktive Populationsmittelwerte, die einer Null-Interaktion entsprechen, sind in Tab. 8.2 enthalten.

Tabelle 8.2: Null-Interaktion: Mittelwerte.

	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>
A <sub>1</sub>	2	8	5
A <sub>2</sub>	4	10	7

Wir wollen dann auf eine Null-Interaktion in der Population erkennen, wenn der erwähnte F-Test bei genügender Teststärke zu einem insignifikanten oder aber praktisch nicht bedeutsamen Ergebnis führt. Ergibt der Test jedoch ein statistisch und praktisch bedeutsames Resultat, gilt es, verschiedene Typen der Interaktion zu unterscheiden.

### 8.5.2 Definition verschiedener Typen der Interaktion

Im Abschnitt 4.3 haben wir ausgeführt, daß eine Variable genau dann einen Störfaktor der Populations- oder auch der Situationsvalidität darstellt, wenn sie mit dem Treatmentfaktor „disordinal interagiert“. Zur Veranschaulichung haben wir dort mit der folgenden Mittelwertstabelle gearbeitet, die wir hier nochmals aufnehmen, ohne die Treatments inhaltlich zu spezifizieren:

Tabelle 8.3: Mittelwerte bei disordinaler Interaktion für Faktor A.

	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>
A <sub>1</sub>	0	5	10
A <sub>2</sub>	2	4	6

Den Daten, die wir als Populationsmittelwerte auffassen wollen, kann entnommen werden, daß das Treatment A<sub>2</sub> nur unter der Modalität B<sub>1</sub> zu höheren Werten auf der AV Y führt; unter den beiden übrigen Modalitäten des Faktors B wirkt A<sub>1</sub> (zahlenmäßig) stärker auf die AV. Die geprüfte wissenschaftliche Hypothese WH, (vgl. Abschnitt 4.3) gilt also nur für die mit B<sub>1</sub> symbolisierte Personengruppe oder Situation oder - allgemeiner ausgedrückt - Faktormodalität; die Populations- oder die Situationsvalidität ist daher eingeschränkt.

Zu einer anderen Interpretation würde man gelangen, wenn die *gleichen* Mittelwerte wie in Tab. 8.3 eine *andere Anordnung* aufweisen würden, etwa die folgende :

Tabelle 8.4: Mittelwerte bei nicht-disordinaler Interaktion bzgl. A.

	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>
A <sub>1</sub>	0	4	6
A <sub>2</sub>	2	5	10

In diesem Beispielfall liegen unter der Bedingung  $A_2$  durchgängig höhere Werte auf der AV Y vor als unter  $A_1$ ; dabei sind allerdings die Unterschiede innerhalb der drei Bedingungen  $B_1$ ,  $B_2$  und  $B_3$  nicht gleich. Die geprüfte  $WH_u$  gilt demnach für alle unter  $B_k$  repräsentierte Personen, Situationen oder Faktormodalitäten, und es liegt daher keine Beeinträchtigung der Populations- oder der Situationsvalidität vor.

Wie man aus diesen Beispielen deutlich erkennen kann, bedeutet die Tatsache, daß eine statistisch signifikante Interaktion vorliegt-hiervon sind wir ausgegangen -, nicht notwendigerweise eine Beeinträchtigung der Populations- oder auch der Situationsvalidität. Diese ist nur dann gegeben, wenn ein ganz bestimmter Typ von Interaktion vorliegt: die bereits mehrfach angesprochene Disordinalität (bezüglich eines Faktors x).

Die Unterscheidung zwischen zwei Typen der Interaktion geht u.W. auf Lindquist (1953, 141 f.) zurück: und Lubin (1961) benutzte die Bezeichnungen „disordinal“ für den in Tab. 8.3 repräsentierten Interaktionstyp und „ordinal“ für die in der Tab. 8.4 enthaltene spezielle Anordnung oder Struktur von Mittelwerten. Die Bedeutung dieser Unterscheidung wurde jedoch nur in wenigen Teilbereichen der Psychologie beachtet, etwa im Rahmen des Konzepts der „Aptitude-Treatment-Interaction“ (ATI) - vgl. Cronbach (1957, 1975), Verreck (1974), Schwarzer & Steinhagen (1975), Treiber & Petermann (1976) sowie Garten (1980); siehe ferner Borich & Godbout (1974), Loftus (1978) und Henning (1978) sowie die oben genannte Literatur.

#### 8.5.2.1 Disordinale Interaktion

Um zu einer formalen Definition der beiden Interaktionstypen gelangen zu können, wollen wir uns etwas näher mit den obigen Tabellen befassen, und zwar zunächst mit der Tabelle 8.3.

In dieser Tabelle drückt sich die Disordinalität der Interaktion darin aus, daß die Differenz der Mittelwerte für A in  $B_1$  ein anderes Vorzeichen („minus“) aufweist als die entsprechende Differenz in  $B_2$  („plus“) und  $B_3$  (ebenfalls „plus“). Im Vergleich zu den Mittelwerten in den beiden letztgenannten Spalten sind die Mittelwerte in Spalte  $B_1$  daher „invertiert“.

Wir bezeichnen im folgenden jeweils diejenigen Mittelwerte als „invertiert“, deren Vorzeichen bei Differenzbildung unterschiedlich sind. Dabei ist es wichtig, daß die Differenzbildung zwischen jeweils „korrespondierenden Mittelwerten“ vorgenommen wird. Unter „korrespondierenden Mittelwerten“ verstehen wir die Mittelwerte innerhalb eines einfachen Haupteffektes in Relation zu den Mittelwerten eines beliebigen anderen einfachen Haupteffektes für den gleichen Faktor. Es korrespondieren in diesem Sinne also allgemein bzgl.

Faktor A die Mittelwerte  $\mu(A_i B_k)$  mit den Mittelwerten  $\mu(A_{i'}, B_k)$ , und für unser Beispiel in Tab. 8.3 ergeben sich damit die folgenden Differenzen zwischen bzgl. A korrespondierenden Mittelwerten:

$$\mu(A_1 B_1) - \mu(A_2 B_1) = 0 - 2 = -2 \text{ (einfacher Haupteffekt A in } B_1)$$

$$\mu(A_1 B_2) - \mu(A_2 B_2) = 5 - 4 = +1 \text{ (einfacher Haupteffekt A in } B_2)$$

$$\mu(A_1 B_3) - \mu(A_2 B_3) = 10 - 6 = +4 \text{ (einfacher Haupteffekt A in } B_3)$$

Überträgt man die (Populations-)Mittelwerte in einen Graphen, wobei auf der Ordinate die Werte der AV und auf der Abszisse derjenige Haupteffekt abgetragen wird, für den der Interaktionstyp *nicht* festgestellt werden soll, so entspricht jeder Inversion von Mittelwerten eine Überkreuzung von Regressionsgeraden als Verbindungslinien zwischen je zwei Mittelwerten. Für das Beispiel aus Tab. 8.3 ergibt sich daher genau eine derartige Überschneidung - vgl. Abb. 8.1.

Will man diese Ausführungen in eine Definition der Disordinalität umsetzen, muß man berücksichtigen, daß wir stets von Populationsmittelwerten ausgegangen sind. In der Stichprobe kann leicht eine Inversion von Zellenmittelwerten  $M_{jk}$  auftreten, die statistisch nicht bedeutsam oder signifikant ist, der also in der Population gleiche Zellenmittelwerte entsprechen.

Nach dieser Überlegung wollen wir in Anlehnung an Bracht & Glass (1968), Bracht (1970) und Plomp (1974) eine disordinale Interaktion in der folgenden Weise definieren:

*Die Hypothese, daß eine Interaktion in der Population disordinal bzgl. eines Haupteffektes X ist, soll angenommen werden, wenn für mindestens zwei Differenzen zwischen korrespondierenden Zellenmittelwerten in der Stichprobe gilt, daß*

- (1) *ihre algebraischen Vorzeichen („plus“ oder „minus“) entgegengesetzt sind, und daß*
- (2) *diese Differenzen statistisch signifikant sind, und daß*
- (3) *die experimentellen Effekte, die mit diesen statistischen Signifikanzen verknüpft sind, „praktisch bedeutsam“ sind, d.h. einen bestimmten Mindestwert nicht unterschreiten (vgl. zum Konzept der „praktischen Bedeutsamkeit (oder Signifikanz)“ Abschnitt 9.2 und Teil 11).*

Wenden wir uns nun der ordinalen Interaktion zu!

### 8.5.2.2 Ordinale Interaktion

Eine ordinale Interaktion bzgl. Faktor ist bei den Mittelwerten in Tab. 8.4 gegeben, die wir wieder als Populationsmittelwerte auffassen wollen. Um fest-

stellen zu können, in welcher Weise sich dieser Interaktionstyp von einer disordinalen Interaktion unterscheidet, bilden wir zunächst die Differenzen korrespondierender Mittelwerte:

$$\mu(A_1B_1) - \mu(A_2B_1) = 0 - 2 = -2$$

$$\mu(A_1B_2) - \mu(A_2B_2) = 4 - 5 = -1$$

$$\mu(A_1B_3) - \mu(A_2B_3) = 6 - 10 = -4$$

Da alle Differenzen das gleiche Vorzeichen aufweisen, können keine Inversionen korrespondierender Mittelwerte vorliegen. Werden also Mittelwerte graphisch veranschaulicht, die einer bzgl. Faktor x ordinalen Interaktion entsprechen, treten keine Überschneidungen von Regressionsgeraden auf, wenn man die Populationsmittelwerte betrachtet - vgl. Abb. 8.3.

Nach diesen Ausführungen definieren wir in Anlehnung an Bracht & Glass (1968), Bracht (1970) und Plomp (1974) eine ordinale Interaktion wie folgt:

*Die Hypothese, daß eine Interaktion in der Population bzgl. eines Haupteffektes X ordinal ist, soll angenommen werden, wenn für alle Differenzen zwischen jeweils korrespondierenden Zellenmittelwerten in der Stichprobe gilt, daß*

- (1) *sie alle das gleiche Vorzeichen aufweisen oder aber höchstens teilweise vom Betrage Null sind, oder daß*
- (2) *diejenigen Differenzen, für die (1) nicht gilt, statistisch nicht signifikant, oder praktisch nicht bedeutsam sind.*

Um also eine statistische Interaktion als „ordinal“ beurteilen zu können, dürfen also keine statistisch abgesicherten Mittelwertsinversionen auftreten.

Wenden wir nun diese Definition an, um zu beurteilen, von welchem Typ die Interaktion bzgl. des bislang nicht betrachteten Faktors B in der Tab. 8.3 ist! Es lassen sich die folgenden drei Gruppen von je zwei korrespondierenden Mittelwertsdifferenzen bilden:

$$\begin{array}{l} \text{I:} \quad \mu(A_1B_1) - \mu(A_1B_2) = 0 - 5 = -5 \\ \quad \mu(A_2B_1) - \mu(A_2B_2) = 2 - 4 = -2 \end{array}$$

$$\begin{array}{l} \text{II:} \quad \mu(A_1B_1) - \mu(A_1B_3) = 0 - 10 = -10 \\ \quad \mu(A_2B_1) - \mu(A_2B_3) = 2 - 6 = -4 \end{array}$$

$$\begin{array}{l} \text{III:} \quad \mu(A_1B_2) - \mu(A_1B_3) = 5 - 10 = -5 \\ \quad \mu(A_2B_2) - \mu(A_2B_3) = 4 - 6 = -2 \end{array}$$

Da in den drei Paaren jeweils korrespondierender Mittelwertsdifferenzen nur jeweils gleiche Vorzeichen auftreten, ist die betr. Interaktion ordinal bzgl. Faktor B.

Die zugehörige graphische Darstellung findet sich in Abb. 8.2, und die Abb. 8.4 enthält die Mittelwerte aus Tab. 8.4, in der die Interaktion bzgl. Faktor B ebenfalls ordinal ist, wie man sich leicht selbst überzeugen kann.

Bredenkamp (1975, 1980) geht von einer anderen Definition der Ordinalität der Interaktion aus, die darauf hinausläuft, daß Null-Differenzen (s.o.) im Sinne der Disordinalität interpretiert werden (vgl. Bredenkamp, 1975, 793, 798). Stanley (1973) unterscheidet die beiden Interaktionstypen nach dem Kriterium ihrer Transformierbarkeit: ordinale Interaktionen sind durch geeignet gewählte Transformationen der Werte der AV eliminierbar, disordinale dagegen nicht - vgl. dazu auch Smith (1976 a) und Busemeyer (1980).

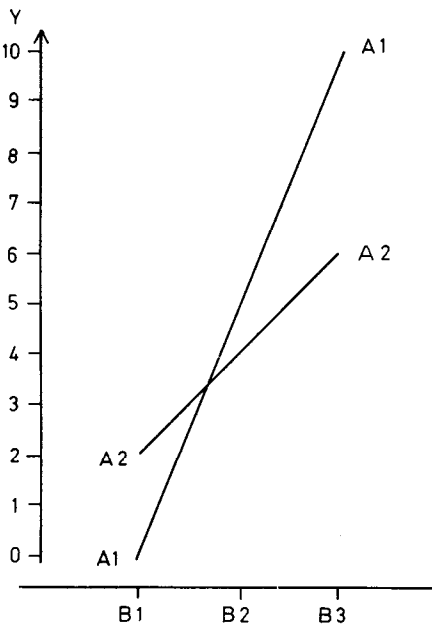


Abb. 8.1: Disordinale Interaktion für Faktor A; Daten aus Tab. 8.3.

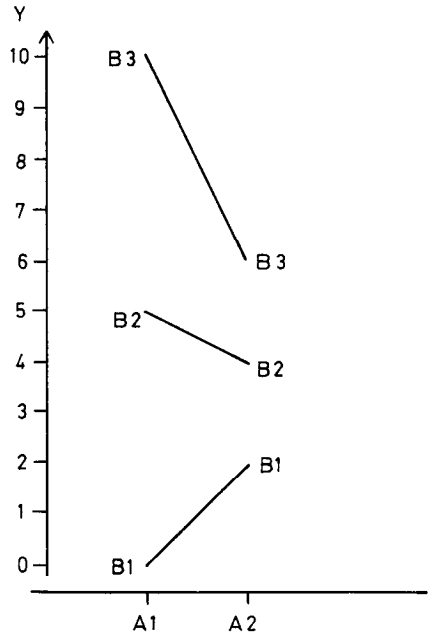


Abb. 8.2: Ordinale Interaktion für Faktor B; Daten aus Tab. 8.3.

### 8.5.2.3 Zur graphischen Darstellung von Interaktionen

Die Definition verschiedener Interaktionstypen abschließend, ist noch eine wichtige Anmerkung angebracht. Häufig wird in der Literatur nicht differenziert, für welchen Faktor eine Interaktion ordinal oder disordinal ist, sondern lediglich allgemein festgestellt, eine Interaktion sei ordinal (oder disordinal). Diese Beurteilung erfolgt oft ausschließlich aufgrund der graphischen Darstellung der Stichprobenzellenmittelwerte. üblicherweise wird dabei jedoch nur einer der beiden möglichen Graphen konstruiert. Tritt dann der Fall auf, daß

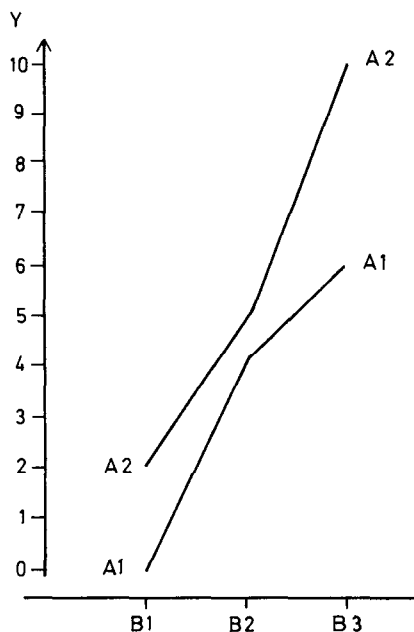


Abb. 8.3: Ordinale Interaktion für Faktor A; Daten aus Tab. 8.4.

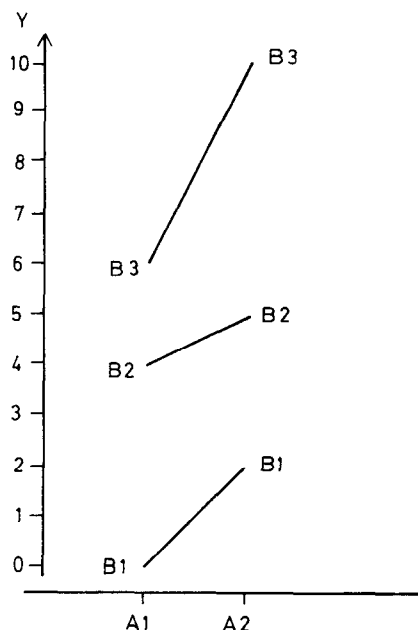


Abb. 8.4: Ordinale Interaktion für Faktor B; Daten aus Tab. 8.4.

die Interaktion für den einen Faktor ordinal, für den anderen aber disordinal ist (vgl. Tab. 8.2), hängt es von der (meist eher zufällig) gewählten Darstellungsform ab, ob sich die Regressionsgeraden schneiden oder nicht (vgl. Abb. 8.1 und 8.2). Durch diese Praxis werden falsche Schlußfolgerungen wie die folgende begünstigt: „It should be emphasized that ‚ordinality‘ and ‚disordinality‘ are properties of graphs“ (Glass & Stanley, 1970, 411); oder Huck & Sutton (1975, 190): „Unfortunately, whether or not an interaction is ordinal can depend upon the way the researcher constructs his graph“; oder auch: „Es muß betont werden, daß ‚Ordinalität‘ und ‚Disordinalität‘ Eigenschaften der graphischen Darstellung von Interaktionen sind“ (McGuigan, 1979, 144).

Selbstverständlich ist der Typ der Interaktion unabhängig von der Art und Weise, in der er graphisch veranschaulicht wird. Zur Vermeidung derartiger Fehlschlüsse und darauf aufbauender Interpretationen bzgl. der Populations- und/oder Situationsvalidität ist es grundsätzlich empfehlenswert, entsprechende Forschungsbefunde auf *beide* Arten darzustellen.



### 8.5.3 *Ein Verfahren zur Unterscheidung zwischen den Interaktionstypen*

Sich überkreuzende Regressionsgeraden, die Stichprobenmittelwerte miteinander verbinden, sprechen nicht notwendig für die Disordinalität der Interaktion in der Population; sie stellen lediglich eine notwendige, nicht jedoch hinreichende Bedingung für die Annahme der Hypothese der Populations-Disordinalität dar. Diese Tatsache führt uns zu der Frage, auf welche Art man über den Typ der Interaktion in der Population aufgrund von Stichprobendaten entscheiden kann.

Zur Beantwortung dieser Frage muß zunächst bedacht werden, daß ein statistisch signifikanter F-Wert für den Test der Interaktion lediglich aussagt, daß die Differenzen zwischen den Zellenmittelwerten in der Population nicht gleich Null sind, daß also in der Population keine Null-Interaktion vorliegt.

Die speziellen Muster oder Strukturen von Mittelwerten, die den verschiedenen Interaktionstypen entsprechen, können mit einem einzelnen F-Wert selbstverständlich nicht erfaßt werden. Dies bedeutet, daß zur Beurteilung des Interaktionstyps i. a. weitere Signifikanztests einzusetzen sind. Um ein allzu starkes Anwachsen der beiden Fehlerwahrscheinlichkeiten zu vermeiden, ist anzustreben, die Zahl dieser Tests so gering wie möglich zu halten. Nach diesem Kriterium sind die verschiedenen in der Literatur vorgeschlagenen mehrstufigen Teststrategien zur Beantwortung der Frage nach dem Interaktionstyp nicht befriedigend (vgl. Bracht & Glass, 1968; Plomp, 1974; Bredenkamp, 1975, 1980).

Als „Übergangslösung“ wollen wir an dieser Stelle eine zwar im Sinne Henning & Muthigs (1979, 205) „quasistatistische“, u.E. aber pragmatische Strategie in ihren Grundzügen kurz skizzieren, die unter Ausnutzung der Variabilität der Daten die Information erbringt, von welchem Typ die Interaktion ist und ggf. auf welche Modalitäten die (Dis-)Ordinalität beschränkt ist.

Nachdem der Test der Interaktion bei genügender Teststärke zu einem signifikanten Ergebnis geführt hat, stellt man zunächst für beide Faktoren getrennt fest, ob irgendwo zwischen jeweils zwei Mittelwertspaaren Inversionen vorliegen. Ist dies für einen oder beide Faktoren nicht der Fall, kann man nach unserer Definition auf Ordinalität der Interaktion für diesen Faktor in der Population erkennen.

Haben sich dagegen Inversionen ergeben, bestimmt man die sog. „minimale statistisch signifikante Differenz (LSD)“ (vgl. zur Berechnung u.a. Winer, 1971, 199f.) und die dem praktisch bedeutsamen experimentellen Mindesteffekt entsprechende „minimale praktisch signifikante Differenz (LPD)“.

Anschließend vergleicht man für jeden einfachen Haupteffekt eines Faktors den maximalen und den minimalen empirischen Mittelwert mit der LSD und der LPD. Alle einfachen Haupteffekte, in denen diese Vergleiche nicht zur Überschreitung der *beiden* kritischen Differenzen führen, bleiben von der weiteren Betrachtung ausgeschlossen, da es in ihnen keine signifikanten Differenzen gibt. Für die übrigen einfachen Haupteffekte wird nun nach Inversionen korrespondierender Mittelwerte gesucht, und jede festgestellte Inversion wird mit der LSD und der LPD verglichen. Ergeben diese Vergleiche für mindestens eine Inversion, daß die beiden Mittelwertsdifferenzen statistisch und praktisch signifikant sind, ist die Interaktion für den untersuchten Faktor disordinal.

Es bleibt anzumerken, daß die hier (u.W. in dieser Form erstmals) vorgestellte Verfahrensweise mit steigender Anzahl der Modalitäten zu anwachsenden Fehlerraten führt. Dies Anwachsen ist jedoch wesentlich geringer als das mit der von Bredenkamp (1975) vorgeschlagenen Teststrategie verbundene; und darüber hinaus wird nur dann auf Disordinalität erkannt, wenn mindestens eine *Konjunktion* von Tests zu einem statistisch und praktisch signifikanten Resultat führt. In Abhängigkeit von den jeweils zu prüfenden WH können sich andere Strategien empfehlen (s. Hager, im Druck, a).

## 8.6 Zusammenfassung

Die Validität einer empirischen Untersuchung zur Prüfung einer wissenschaftlichen Hypothese kann u. a. dadurch vermindert werden, daß man zu falschen Aussagen über die Gültigkeit der aus der wissenschaftlichen Hypothese abgeleiteten statistischen Hypothese gelangt. Die maximalen Wahrscheinlichkeiten  $\alpha$  und  $\beta$  für die beiden Fehlentscheidungen über die Gültigkeit einer statistischen Hypothese sollten nach unseren Überlegungen in den Teilen 6 und 7 durch den Experimentator von vornherein auf möglichst niedrige Werte fixiert werden. Ungeachtet dieser Festlegung kann es jedoch aufgrund von Fehlern bei der Prüfung statistischer Hypothesen dazu kommen, daß sich die tatsächlichen Wahrscheinlichkeiten für falsche Entscheidungen erhöhen; diese Fehler haben wir als „Störfaktoren der statistischen Validität“ bezeichnet. Die statistische Validität von Untersuchungen wird häufig dadurch gestört, daß anstelle der von der wissenschaftlichen Hypothese implizierten eine ganz andere statistische Hypothese geprüft wird. Insbesondere müßten in der Psychologie statt der verbreiteten Varianz- und Regressionsanalysen viel öfter spezielle Mittelwertvergleiche wie die Prüfung monotoner Trendhypothesen angestellt werden, zumindest im Anschluß an signifikante globale Tests (Abschnitt 8.1 und 8.2). Um eine Kumulierung von Fehlerwahrscheinlichkeiten weitestgehend zu vermeiden, ist in Abhängigkeit von der geprüften wissenschaftlichen Hypo-

these entweder die Wahrscheinlichkeit  $\alpha$  oder aber  $\beta$  für jeden einzelnen Test gegenüber den „konventionellen“ Werten herabzusetzen.

Für die abgeleitete statistische Hypothese muß dann ein angemessenes statistisches Prüfverfahren gewählt werden. Sind die Annahmen der klassischen parametrischen Tests (vgl. Abschnitt 8.2.1) nicht erfüllt, kann es zu einer Erhöhung der Fehler-Wahrscheinlichkeiten kommen, die allerdings i. a. nur unter besonders ungünstigen Umständen erhebliche Ausmaße annehmen dürften (Abschnitt 8.2). Muß man davon ausgehen, daß zwei oder mehr Annahmen nicht aufrechterhalten werden können, sollte man die Daten i. a. nicht mittels eines parametrischen Verfahrens auswerten.

Im Abschnitt 8.4 haben wir einige Methoden angesprochen, die die Varianz der abhängigen Variablen innerhalb der Behandlungspopulationen verringern und dadurch die Präzision der statistischen Hypothesenprüfung erhöhen können: die Parallelisierung der Untersuchungseinheiten (Vpn) hinsichtlich individueller Merkmale, die Kovarianzanalyse, die Beobachtung aller Vpn unter mehreren (oder allen) Modalitäten eines Faktors (Meßwiederholung), die Verwendung homogener Probandengruppen sowie die Konstanthaltung und Elimination situationaler Variablen. Bei der Anwendung dieser Techniken ist jedoch zweierlei zu beachten: Erstens führen gerade die bekanntesten von ihnen (Parallelisierung und Meßwiederholung) nur unter ganz bestimmten Bedingungen tatsächlich zu einer erhöhten Präzision, und zweitens kann der Einsatz dieser Techniken andere Aspekte der experimentellen Validität beeinträchtigen. Bei Meßwiederholungen ist bspw. die interne Validität meist nur unzureichend zu sichern, während Homogenisierung, Konstanthaltung und Elimination leicht zur Beeinträchtigung der Populations- und/oder der Situationsvalidität führen können.

Im Abschnitt 8.5 haben wir auf eine häufig vernachlässigte Störung der statistischen Validität hingewiesen: die falsche Analyse und Interpretation statistischer Interaktionen. Um beurteilen zu können, ob das Vorliegen einer Interaktion die Interpretation der Ergebnisse statistischer Tests über Mittelwerte (oder Ränge) beeinflusst, muß festgestellt werden, von welchem Typ diese Interaktion ist. Die Disordinalität der Interaktion führt zu Einschränkungen der Interpretation hinsichtlich der Populations- oder der Situationsvalidität, während dies bei Vorliegen einer ordinalen Interaktion nicht der Fall ist. Um den Typ einer Interaktion bzgl. eines Faktors mit einem Minimum an statistischen Tests feststellen zu können, wurde ausgehend von einer Definition der Disordinalität eine neue Strategie vorgestellt.

## 9. Maße der statistischen Assoziation: Die experimentellen Effekte

### 9.1 Einleitung

In den vorangegangenen Teilen haben wir jegliche Abweichung eines Parameters von seinem Erwartungswert unter Gültigkeit der statistischen Null-Hypothese mit „experimenteller Effekt (EE)“ bezeichnet.

Der Terminus „experimenteller Effekt“ ist demnach ein allgemeiner Ausdruck für systematische Beziehungen, d.h. die statistische Assoziation zwischen den jeweils untersuchten Variablen; in einem Experiment bspw. entsteht diese Assoziation zwischen UV(n) und AV(n) infolge der experimentellen Behandlungen.

Bei der Bestimmung des EE mittels Stichprobendaten muß stets damit gerechnet werden, daß unsystematische Fehler zu Verzerrungen führen, d.h. im Extremfall entweder zu „Scheineffekten“ oder aber zu „fälschlichen Null-Effekten“.

Der Signifikanztest dient unter diesem Aspekt der Beantwortung der Frage, ob ein aus den Daten ermittelter Effekt als durch die unsystematischen Fehler zustande gekommen erklärt werden kann oder ob er eine tatsächlich vorhandene systematische Beziehung zwischen UV und AV reflektiert. Unter sonst gleichen Bedingungen wird ein EE um so eher als „systematisch“ oder „echt“ deklariert, je mehr Vpn zu seiner Entdeckung verwendet wurden.

Da das Vorhandensein eines EE von der Tatsache der statistischen Signifikanz unabhängig ist, sollte seine Größe ungeachtet des Ausganges des Signifikanztests stets aus den erhobenen Daten ermittelt werden (vgl. Abschnitt 7.4).

Spätestens hierbei stellt sich die Frage nach der (formalen) Definition von experimentellen Effekten.

Grundsätzlich sind EE stets unter Verwendung derjenigen Größen definiert, über die die statistischen Hypothesen formuliert werden - also etwa über Mittelwerte, Korrelationskoeffizienten, Ränge oder Wahrscheinlichkeiten. Auf diese Definitionen gehen wir im folgenden detaillierter ein, wobei wir erneut die auf Intervall-Informationen beruhenden EE ausführlicher darstellen werden.

Zuvor jedoch ist eine grundsätzliche Unterscheidung anzusprechen.

## 9.2 Experimentelle Effekte und praktische Bedeutsamkeit

Die EE wurden in der psychologischen Versuchsplanungsliteratur u.W. von Edwards (1950, 30f.) unter dem Namen „practical significance“ eingeführt und finden sich unter wechselnden Bezeichnungen bei anderen Autoren. Hodges & Lehmann (1954, 261) sprechen etwa von „material significance“, Bolles & Messick (1959) von „statistical utility“, während Kendall & Stuart (1961, 161) die Bezeichnung „magnitude of a difference“ wählen. Hays (1963, 326) nennt den EE „degree (oder auch: „strength“) of true association“; Levy (1967) und Gold (1969) bezeichnen den EE als „Substantive significance“; Vaughan & Corballis (1969) reden von „strength of effect“ und Cohen (1969) kurz von „effect size“. Für den deutschen Sprachraum haben offenbar Lienert & Orlik (1966, 215) die Bezeichnung „praktische Signifikanz“ geprägt, die Bredenkamp (e. g. 1970) später aufgegriffen hat.

Wir plädieren dafür, die Begriffe „experimenteller Effekt“ einerseits und „praktische Bedeutsamkeit“ (oder: „praktische Signifikanz“) andererseits voneinander zu trennen.

Unter „experimenteller Effekt“ (EE) verstehen wir die rein statistische Assoziation, die sich in den Daten manifestiert bzw. aus diesen zu bestimmen ist und die entweder in Roheinheiten (etwa als Mittelwertsdifferenz) oder in normierten Einheiten (etwa als Korrelationskoeffizient; s.U.) angegeben werden kann. Der Ausdruck „experimenteller Effekt“ bezeichnet damit ein ausschließlich statistisches Konzept und bezieht sich auf das Ausmaß der Abweichung eines Parameters von der Erwartung unter einer einfachen Null-Hypothese. Ob man diese Abweichung im konkreten Einzelfall inhaltlich als „praktisch bedeutsam“ im Hinblick auf die zu prüfende wissenschaftliche Hypothese interpretieren will, ist eine Entscheidung, für die es ein statistisches Rationales u.W. nicht gibt, für die also der E andere Kriterien angeben sollte - vgl. zu Einzelheiten Teil 11.

Wir befassen uns im Teil 9 ausschließlich mit den Maßen für die statistische Assoziation bzw. für den experimentellen Effekt in der Population und in der Stichprobe und greifen das Konzept der praktischen Bedeutsamkeit erst im Teil 11 wieder auf.

## 9.3 Experimentelle Effekte bei parametrischen Hypothesen

Parametrische Hypothesen beziehen sich in der überwiegenden Mehrzahl der Fälle auf Mittelwerte bzw. Regressionskoeffizienten und quadrierte Korrelationen. Die Prüfung dieser Hypothesen erfolgt in aller Regel über die Teststatistiken F und t.

Im Abschnitt 7.3.2 haben wir festgestellt, daß diese Statistiken bei Vorhandensein eines experimentellen Effektes in der Population, EEP, nicht-zentralen Verteilungen folgen. Das Ausmaß der Nicht-Zentralität, d.h. der Abweichung von den zentralen Verteilungen, wird durch die Nicht-Zentralitätsparameter erfaßt. Diese enthalten also Informationen über den EEP.

### 9.3.1 Maße der Nicht-Zentralität: $\lambda$ , $\varphi^2$ und $f^2$

Die in diesem Abschnitt vorzustellenden Maße der Nicht-Zentralität sind allesamt unter Verwendung der quadrierten Differenz der K Treatment-Mittelwerte  $\mu_k$  von ihrem Gesamtmittelwert  $\mu$  definiert; trifft  $H_0$  zu, müssen alle Mittelwerte  $\mu_k$  identisch sein. Je größer also diese (quadrierten) Abweichungen sind, desto größer ist auch unter sonst gleichen Bedingungen der EEP.

Ferner ist allen Nicht-Zentralitätsparametern gemeinsam, daß die quadrierten Mittelwertsabweichungen auf die Populationsfehlervarianz  $\sigma_e^2$  normiert werden, also als Abweichungen relativ zum Fehler ausgedrückt werden. Durch diese Normierung wird die Vergleichbarkeit der in möglicherweise sehr unterschiedlichen Experimenten ermittelten Abweichungen angestrebt, die eine Voraussetzung zur Tabellierung der Gütefunktionen darstellt (vgl. Abschnitt 7.4).

Unter der Annahme gleicher Größe der K (Sub-)Populationen ergibt sich folgende allgemeine Definition der Nicht-Zentralitätsparameter:

$$(9.1) \quad \text{Nicht-Zentralität:} = c \cdot \frac{\sum_{k=1}^K (\mu_k - \mu)^2 / K}{\sigma_e^2} = c \cdot \frac{\sigma_b^2}{\sigma_e^2}$$

Die in der Literatur auffindbaren Nicht-Zentralitätsparameter, die sehr uneinheitlich symbolisiert werden, unterscheiden sich im wesentlichen durch die Größe, die an die Stelle der (allgemeinen) Konstanten  $c$  in (9.1) tritt. Der Parameter  $\lambda$ , zuweilen auch als  $\delta^2$  bezeichnet, ergibt sich, indem man  $c = n \cdot K$  setzt, wobei  $n$  die Stichprobengröße pro experimenteller Bedingung und  $K$  die Anzahl dieser Bedingungen bezeichnet (vgl. etwa Kirk, 1968, 107) (wir gehen hier wie im folgenden stets von gleichen Stichprobenumfängen aus):

$$(9.2) \quad \lambda = \frac{n \cdot K \cdot \sigma_b^2}{\sigma_e^2} = \frac{N \cdot \sigma_b^2}{\sigma_e^2}$$

Tang (1938), der sich als erster mit der Tabellierung der nicht-zentralen F-Verteilungen befaßt hat, benutzte einen aus  $\lambda$  abgeleiteten Index  $\varphi$ , der der Quotient aus der Standardabweichung der wahren Mittelwertsdifferenzen zu

Lasten des experimentellen Treatments und dem Standardfehler des Mittelwerts ist (vgl. Tang, 1938, 138); für das Quadrat dieses Index' gilt daher bei c n:

$$(9.3) \quad \varphi^2 = \frac{n \sum_{k=1}^K (\mu_k - \mu)^2 / K}{\sigma_e^2} = \frac{n \cdot \sigma_b^2}{\sigma_e^2}$$

Einen weiteren Index für die Nicht-Zentralität hat Cohen (1969, 267-280) vorgeschlagen. Sein Maß  $f^2$  erlaubt die Beschreibung einer Populationssituation unabhängig von der Stichprobengröße n bei c = 1:

$$(9.4) \quad f^2 = \frac{\sum_{k=1}^K (\mu_k - \mu)^2 / K}{\sigma_e^2} = \frac{\sigma_b^2}{\sigma_e^2}$$

Zwischen den drei vorgestellten Nicht-Zentralitätsparametern bestehen folgende wichtige Beziehungen, aus denen sich weitere Zusammenhänge durch einfache algebraische Umformungen ableiten lassen:

$$(9.5) \quad f^2 = \frac{\varphi^2}{n} = \frac{\lambda}{n \cdot K}$$

Neben diesen drei Nicht-Zentralitätsparametern als normierte EEP finden sich in der Literatur weitere, die meist auf die hier angegebenen zurückgeführt werden können - vgl. zu Einzelheiten etwa Winer (1971, 826f.). Bei der Beschäftigung mit diesen Maßen sollte stets beachtet werden, daß die Symbole sehr uneinheitlich verwendet werden.

### 9.3.2 Korrelationskoeffizienten und -quotienten

Die Nicht-Zentralitätsparameter können als Funktionen interpretiert werden, die die Stärke eines statistischen Zusammenhangs auf den Bereich der positiven reellen Zahlen von 0 bis  $+\infty$  abbilden. Diese Art der Abbildung ist insgesamt allerdings weniger gebräuchlich als diejenige in den Bereich der positiven reellen Zahlen von 0 bis 1; hierfür ist etwa der Korrelationskoeffizient r (als normierte Kovarianz) ein Beispiel.

Bei dieser Normierungsart wird die Varianz zu Lasten der verschiedenen Treatments, also  $\text{Var}(b)$ , nicht mehr auf die Fehlervarianz  $\text{Var}(e)$  relativiert, sondern auf die totale Varianz in der AV, nämlich  $\text{Var}(t)$ , die sich bei Zugrundelegung des ALM (vgl. Abschnitt 8.2) additiv aus den beiden anderen Varianzen

ergibt (vgl. zu den Ableitungen im einzelnen etwa Hays, 1977, Kap. 12). Man nennt die entstehenden Quotienten „Maße der „erklärten“ oder „aufgeklärten“ Varianz“ (Bredenkamp, 1970; zur Kritik dieser Bezeichnung vgl. etwa Guttman, 1977):

$$(9.6) \quad \text{„erklärte“ Varianz:} = \frac{\text{Var}(t) - \text{Var}(e)}{\text{Var}(t)}$$

In Abhängigkeit davon, welche Varianzen man in die allgemeine Formel (9.6) einsetzt, erhält man normierte Maße für den EEP oder für den EES, den experimentellen Effekt in der Stichprobe.

### 9.3.2.1 Populationsmaße: $\eta^2$ , $\omega^2$ und $R_{Y.X}^2$

Ersetzt man die allgemeinen Varianzterme in (9.6) durch die entsprechenden Populationsvarianzen, erhält man:

$$(9.7) \quad \eta^2 = \frac{\sigma_t^2 - \sigma_e^2}{\sigma_t^2} = \frac{\sigma_b^2}{\sigma_t^2} = 1 - \frac{\sigma_e^2}{\sigma_t^2}$$

Diese Quotientenbildung geht auf Pearson (e. g. 1911) zurück und das Resultat wird nach Fisher (1928) in der englisch-sprachigen Literatur „correlation ratio“ genannt. Im deutschen Sprachraum finden sich austauschbar die Bezeichnungen „Korrelationsverhältnis“ oder „Korrelationsindex“ (vgl. etwa Lienert & Raatz, 1971) oder auch „Korrelationsquotient“ und „Eta-Quotient“ (Kerlinger, 1978, 336; 1979, 1042).

$\eta^2$  gibt den relativen Anteil der Gesamtvariation in der AV an, der auf die Variation der UV zurückgeführt werden kann, also „systematisch“ ist; es beschreibt jegliche funktionale Beziehung zwischen einer beliebigen UV und einer kontinuierlichen AV - weitere Interpretationen dieses Maßes finden sich etwa bei Cohen (1965, 104f.). Liegt eine quantitativ gestufte (also metrische) UV vor,<sup>31)</sup> ist es sinnvoll,  $\eta^2$  als ein Maß aufzufassen, das - im Gegensatz zum bekannteren Produkt-Moment-Korrelationskoeffizienten  $r$  - den linearen und alle kurvilinearen Zusammenhänge zwischen der UV und der AV zu erfassen erlaubt; zu weiteren Einzelheiten siehe Cohen (1965, 104f.) und Hays (1977, 682-684).

---

<sup>31)</sup> Als Beispiel denke man an die Geldbeträge, die ein die Theorie der kognitiven Dissonanz prüfender E aussetzt, um seine Vpn zu einer ihrer Meinung widersprechenden Aussage zu veranlassen (Festinger & Carlsmith, 1959), oder aber an Darbietungszeiten in einem Lernexperiment (Bredenkamp & Hager, 1979); vgl. auch Abschnitt 1.1.



Bekannter als  $\eta^2$  ist in den letzten Jahren das von Hays (1963, 325, 382; 1977, 414) vorgeschlagene  $\omega^2$  geworden:

$$(9.8) \quad \omega^2 = \frac{\sigma_t^2 - \sigma_e^2}{\sigma_t^2}$$

Verläßt man das Modell der Varianzanalyse mit fixierten Effekten und geht auf das der multiplen Regression über, läßt sich den einschlägigen Lehrbüchern (etwa Kerlinger & Pedhazur, 1973; Cohen & Cohen, 1975) entnehmen, daß noch ein weiteres Maß verwendet werden kann, um die statistische Assoziation in der Population anzugeben, nämlich  $R_{Y.X}^2$ , das multiple Korrelationsquadrat. Kodiert man die UV bzw. die Gruppenzugehörigkeit mittels sog. Dummy-Variablen und setzt man voraus, daß den einzelnen Bedingungen formal „Zellenwahrscheinlichkeiten“ zugewiesen werden können (Keren & Lewis, 1979), gilt in der Population die numerische Äquivalenz von  $R_{Y.X}^2$ ,  $\omega^2$  und  $\eta^2$  (Cohen & Cohen, 1975, 187f.; Cohen, 1977).

Um auch in mehrfaktoriellen Plänen die experimentellen Effekte zu Lasten der einzelnen Faktoren bestimmen zu können, definiert man das Maß für den normierten EEP zweckmäßigerweise als quadrierte Partialkorrelation. Hierdurch wird der relative Anteil der Effektvarianz zu Lasten bspw. des Faktors A, also  $\sigma_A^2$ , an der totalen Varianz  $\sigma_t^2$  erfaßt, aus der sämtliche Variationsquellen zu Lasten der übrigen Faktoren und der Interaktion(en) auspartialisiert werden - vgl. im einzelnen Kennedy (1970), Cohen (1973a, 1977), Humphreys & Fleishman (1974), Cohen & Cohen (1975) sowie Keren & Lewis (1979). Somit ergibt sich bspw.:

$$(9.9) \quad R_{YA.B,AB,\dots}^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}$$

Wie sich den Definitionen entnehmen und durch algebraische Umformungen zeigen läßt, gelten folgende Beziehungen zwischen den hier angesprochenen Korrelationskoeffizienten und den Nicht-Zentralitätsparametern im vorigen Abschnitt (vgl. auch Cohen, 1977, 410-415, und Hager, im Druck, b):

$$(9.10) \quad f^2 = \frac{R_{Y.X}^2}{1 - R_{Y.X}^2}$$

$$(9.11) \quad R_{Y.X}^2 = \frac{f^2}{1 + f^2}$$

Plant man ein multivariates Experiment zur Prüfung von Hypothesen über Mittelwertsvektoren, müssen die soeben vorgestellten univariaten Maße für den EE durch ihre multivariaten Äquivalente ersetzt werden, die nach den gleichen Prinzipien definiert sind wie die univariaten Maße; nähere Einzelheiten entnehme man den Arbeiten von Cooley & Lohnes (1971, 227), Huberty

(1972), Smith (1972), Stevens (1972a, 1980), Sachdeva (1973), Olson (1974), Shaffer & Gillo (1974), Bredenkamp (1975, 1980), Odeh & Fox (1975, 36-42), Läuter (1978) sowie Cramer & Nicewander (1979).

### 9.3.2.2 Stichprobenmaße: $\hat{R}_{Y.X}^2$ , $E^2$ , UI

Ersetzt man in der Ausgangsdefinition (9.6) die Varianzterme durch die Stichprobenvarianzen  $S_t^2$  und  $S_e^2$ , erhält man normierte Maße für den EES, die als Stichprobenäquivalente der entsprechenden Populationsmaße aufgefaßt werden können. Auch im Stichprobenfall wird dieser Quotient mit zahlreichen unterschiedlichen Symbolen belegt; Wishart (1932) und Kerlinger (1964) bezeichnen ihn mit  $E^2$ , andere Autoren, etwa Cohen (1965, 105), mit  $\eta^2$ . Bolles & Messick (1958) und Gaito (1958) haben den Namen „Utilitätsindex“ UI oder  $U_b$  vorgeschlagen; dieser Vorschlag wurde neuerdings von Gaebelin & Soderquist (1978) und Soderquist & Hussian (1978) wieder aufgegriffen.

Aus der folgenden Definitionsgleichung (9.13) wird jedoch deutlich, daß die unterschiedlichen Symbole allesamt die quadrierte multiple Korrelation in der Stichprobe,  $\hat{R}_{Y.X}^2$ , bezeichnen - vgl. Kerlinger & Pedhazur (1973) sowie Cohen & Cohen (1975):

$$(9.12) \quad \hat{R}_{Y.X}^2 = \frac{S_t^2 - S_e^2}{S_t^2} = \frac{S_b^2}{S_t^2} = 1 - \frac{S_e^2}{S_t^2} = \frac{SST - SSE}{SST}$$

(„SST“ bezeichnet die varianzanalytische totale Quadratsumme, „SSE“ die Fehlerquadratsumme.)

Werden nur  $K = 2$  Mittelwerte untersucht, ist der Ausdruck (9.12) mit dem quadrierten Punkt-biserialen Korrelationskoeffizienten  $r_{pbis}^2$  identisch (Cohen, 1968; Bredenkamp, 1970).

Darüber hinaus ist  $\hat{R}_{Y.X}^2$  nichts anderes als eine monotone Transformation der empirischen Werte der üblichen Teststatistiken  $F$  und  $t$ , wie bereits Fisher (1928, 672) und Wishart (1932, 441) gezeigt haben; und es gilt ( $F$  bezeichne den empirischen Wert der Teststatistik  $F$ ):

$$(9.13) \quad \hat{R}_{Y.X}^2 = \frac{(K - 1) F}{(K - 1) F + (N - K)}$$

Weitere Maße für den EES haben Levy (1967; vgl. auch Cohen, 1969, 19-21) und Friedman (1968, 1969; zur Kritik Breen & Gaito, 1970) vorgeschlagen, die mit den hier angesprochenen in sehr engem Zusammenhang stehen und deshalb nicht gesondert erörtert werden - vgl. zu Einzelheiten Bredenkamp (1970) sowie Gaito & Firth (1973).

### 9.3.2.3 Korrekturformeln für $\hat{R}_{Y.X}^2$

Es ist bekannt, daß die Größe der multiplen Korrelation in der Stichprobe,  $\hat{R}_{Y.X}$ , unter sonst gleichen Bedingungen um so größer ist, je mehr Modalitäten der UV X - also „Prädiktoren“ - zur Vorhersage des „Kriteriums“ - also der Werte der AV - verwendet werden. Selbst wenn die tatsächliche Korrelation zwischen den Prädiktoren und dem Kriterium gleich Null ist, ist im Mittel über alle möglichen Stichproben ein Korrelationsquadrat der Größe

$$(9.14) \quad \hat{R}_{Y.X}^2 = \frac{K - 1}{N - 1}$$

zu erwarten (vgl. Wishart, 1932, 444; Cohen & Cohen, 1975, 107). Zur Reduktion dieser systematischen Verzerrung („Bias“) werden in der Literatur unterschiedliche sog. „(Schrumpfungs-)Korrekturen“ vorgeschlagen, über die zusammenfassend Särndal (1974), Carter (1979) und Huberty & Mourad (1980) informieren.

Von diesen Korrekturformeln greifen wir zwei heraus, die häufiger im Zusammenhang mit Varianz- und Regressionsanalysen benutzt werden.

$$(9.15) \quad \hat{R}_{\text{kor.}}^2 = \epsilon^2 = 1 - (1 - \hat{R}^2) \left( \frac{N - 1}{N - K} \right) = \frac{(K - 1)(F - 1)}{(K - 1)F + (N - K)} =$$

$$= \frac{s_t^2 - s_e^2}{s_t^2}$$

Hierbei gilt:  $s_t^2 = [(SST) : (N-1)]$  und  $s_e^2 = [(SSE) : (N-K)]$  - siehe zu dieser Formel u.a. Cohen (1965, 105) und Fleiss (1969). Letzterer gibt darüber hinaus eine weitere Korrekturformel an:

$$(9.16) \quad \hat{R}_{\text{kor.}}^2 = \hat{\omega}^2 = \hat{\eta}^2 = 1 - (1 - \hat{R}^2) \left( \frac{N}{N - K + 1 - \hat{R}^2} \right) =$$

$$= \frac{(K - 1)(F - 1)}{(K - 1)(F - 1) + N} = \frac{s_t^2 - s_e^2}{s_t^2 + \frac{s_e^2}{N - 1}}$$

Bredenkamp (1970) hat im einzelnen zeigen können, daß diese unterschiedlichen Korrekturformeln entstehen, indem jeweils verschiedene Schätzwerte für die Populationsvarianzen  $\sigma_t^2$ ,  $\sigma_b^2$  und  $\sigma_e^2$  in die Gleichungen (9.5) und (9.6) eingesetzt werden.

Die Kriterien und Methoden zur Ermittlung dieser Schätzwerte, die Wang (1967) und Miller (1979) miteinander verglichen haben, finden sich in einschlägigen Lehrbüchern wie etwa denen von Winer (1962, 1971), Glass & Stanley (1970), Edwards (1971, 1980), Lee (1975), Bortz (1979) sowie Henning & Muthig (1979) und darüber hinaus in den

Arbeiten von Gaito (1960a), Halderson & Glasnapp (1972), Dwyer (1974), Gaebelein & Soderquist (1976) und Hopkins (1976); Schätzwerte für in einigen sog. „Standard-Versuchsplänen“ zu erwartende Varianzen haben zudem Gaito & Turner (1963), Vaughan & Corballis (1969), Fleiss (1969) sowie Dodd & Schultz (1973) veröffentlicht.

Die o. gen. Korrekturen, die als Schätzungen der Populationskorrelation  $R_{Y,X}^2$  interpretiert werden können, genügen allesamt nicht dem Kriterium der „Erwartungstreue“ (vgl. hierzu etwa Stilson, 1966, 205f.; Hays, 1977, 272-274; Bortz, 1979, 123f.), obwohl dies verschiedentlich behauptet worden ist. Schätzungen, die diesem Kriterium genügen, liegen bislang nicht vor - vgl. dazu Olkin & Pratt (1958), Cureton (1966), Bredenkamp (1970) und Huberty & Mourad (1980).

Zwar meinen wir, daß grundsätzlich korrigierte Werte für  $\hat{R}_{Y,X}^2$  den unkorrigierten vorzuziehen sind, aber andererseits kann es dem E im Einzelfall überlassen bleiben, für welche der möglichen Korrekturen er sich entscheidet, weil die Unterschiede insgesamt Vernachlässigbar gering sind und bei größeren Stichprobenumfängen ohnehin verschwinden; vgl. hierzu die Computer-Simulationen („Monte-Carlo-Studien“)<sup>32)</sup> von Särndal (1974), Carroll & Nordholm (1975), Keselman (1975), Carter (1979) sowie von Huberty & Mourad (1980).

Diesen Untersuchungen ist jedoch auch zu entnehmen, daß die Standardfehler der Korrelationsstatistiken meist unverhältnismäßig hoch sind. Da die theoretische Ableitung der entsprechenden Stichprobenverteilungen bislang noch nicht gelungen ist, kann im Einzelfall nicht angegeben werden, wie stark die empirischen Werte für  $\hat{R}_{Y,X}^2$  oder  $\hat{R}_{\text{kor.}}^2$  streuen können (zur Bestimmung approximativer Konfidenzintervalle siehe Venables, 1975, und Fleishman, 1980).

Allerdings sind teilweise für Nicht-Zentralitätsparameter wie etwa  $f^2$  (vgl. Abschnitt 9.3.1) die Standardfehler bestimmt worden. Man kann also den Wert für das korrigierte multiple Korrelationsquadrat in  $\hat{f}^2$  umrechnen und für dieses den Standardfehler errechnen - vgl. zu Einzelheiten Fleishman (1980).

Zur direkten Bestimmung des Ausmaßes der Nicht-Zentralität in der Stichprobe als monotone Funktion des multiplen Korrelationsquadrats oder des empirischen Wertes der Teststatistik  $F$  (resp.  $t$ ) können die Gleichungen (9.10) und (9.15) oder (9.16) herangezogen werden.

Darüber hinaus kann der Nachteil des (für Korrelationskoeffizienten typischen) großen Standardfehlers bei der Beurteilung der WH teilweise dadurch ausgeglichen werden, daß die EES der entsprechenden Experimente zusammenfassend miteinander verglichen werden - siehe Abschnitt 11.2.

---

<sup>32)</sup> Zur Einführung in diese Techniken orientiere man sich u.a. bei Hammersley & Handscomb (1964), Collier & Larson (1969), Kleijnen (1974, 1975) sowie Bauknecht, Kohlas & Zehnder (1976).

## 9.4 Experimentelle Effekte bei nicht-parametrischen Hypothesen

Auch bei ordinalen und nominalen Daten sind die EE in Abhängigkeit von den jeweils geprüften Hypothesen zu definieren; dies führt zu einer Fülle von unterschiedlichen Maßen für den normierten EE. Wir wollen im folgenden nur einige gebräuchliche Konstruktionsprinzipien und Maße aufgreifen.

### 9.4.1 Experimentelle Effekte bei ordinalen Daten

Das nicht-parametrische Äquivalent der (auf Intervallniveau beruhenden) parametrischen Korrelationskoeffizienten stellen die Rangkorrelationskoeffizienten dar.

Liegen bspw. in einem Zwei-Gruppen-Plan abhängige rangskalierte Daten vor, kann der Zusammenhang zwischen beiden Rangreihen über die rechnerische Vereinfachung des Produkt-Moment-Korrelationskoeffizienten  $r$  bestimmt werden, die in der Literatur üblicherweise „Spearman's Rangkorrelationskoeffizient  $r_s$ “ genannt wird (vgl. etwa Bortz, 1979, 283-286; ferner Hays, 1977, 788-792).

Für einen Versuchsplan mit  $K$  Gruppen und unabhängigen Messungen haben Lienert & Raatz (1971) das nicht-parametrische Äquivalent des parametrischen multiplen Korrelationsquadrates abgeleitet, das sie „Rangkorrelationsverhältnis  $\eta_H^2$ “ nennen; wir wollen im folgenden die Bezeichnung  $E_H^2$  benutzen, um anzudeuten, daß es sich um ein Maß für den EE in der Stichprobe handelt. Dies folgt aus der Definition von  $E_H^2$  als Funktion des empirischen Wertes der Teststatistik  $H$  (vgl. dazu Kruskal & Wallis, 1952):

$$(9.17) \quad E_H^2 = \frac{H}{N - 1}$$

Der Quotient (9.17) gibt den relativen Anteil an der gesamten Rangvarianz an, der durch die Treatments „aufgeklärt“ werden konnte, und entspricht konzeptuell dem parametrischen  $\hat{R}_{Y.X}^2$  (vgl. Lienert & Raatz, 1971; Bredenkamp, 1980, 62).

Die bisher vorgestellten Normierungen des in absoluten Einheiten angegebenen EE erfolgten entweder auf die Fehler- oder aber auf die totale Varianz der AV, und es resultierten Maße der Nichtzentralität und der erklärten Varianz (quadrierte Korrelationen). Eine weitere Normierungsmöglichkeit besteht darin, den empirischen Wert einer Teststatistik auf den jeweils theoretisch maximal möglichen Wert zu relativieren. Das Resultat kann man als „Kontingenz-Quotient“ bezeichnen, da der in Kontingenztafeln bestimmbare EE i.a. nach dieser Methode normiert wird (vgl. Abschnitt 9.4.2).

$$(9.18) \quad \text{EE als Kontingenzquotient: } = \frac{\text{ermittelter Wert der Teststatistik}}{\text{maximal möglicher Wert der Teststatistik}}$$

Acock & Stavig (1979; vgl. auch Stavig & Acock, 1980) haben auf dieser Basis für etliche nicht-parametrische Testverfahren Maße für den normierten EE abgeleitet. Als Beispiel ergibt sich für die Rangvarianzanalyse nach Friedman (1937) über abhängige Rangdaten ein Quotient, der dem Konkordanzkoeffizienten  $W$  von Kendall (1970) entspricht.

Die als Kontingenzquotienten definierten Maße für den normierten EE sind allesamt mit den quadrierten Korrelationskoeffizienten (s. o.) nicht vergleichbar, obwohl sie sich ebenfalls innerhalb der Grenzen 0 und 1 bewegen (Lienert & Raatz, 1971).

Weitere Maße des statistischen Zusammenhangs zwischen ordinalen Variablen können der im Abschnitt 7.5.3.1 angeführten Literatur entnommen werden; siehe ferner Särndal (1974) und Glass (1978).

#### 9.4.2 Experimentelle Effekte bei nominalen Daten

Das Ausmaß der statistischen Assoziation in einer Vier-Felder-Kontingenztafel wird häufig mittels des „Punkt-Vier-Felder-Korrelationskoeffizienten“, auch kurz „Phi-Koeffizient“ genannt, erfaßt, der ebenfalls eine rechnerische Vereinfachung des Korrelationskoeffizienten  $r$  darstellt.

Wenn wir vom empirischen Wert der Teststatistik  $\chi^2$  ausgehen, ist der quadrierte Phi-Koeffizient wie folgt definiert (Hays, 1977, 743):

$$(9.19) \quad V_P^2 = \frac{\chi^2}{N}$$

Bei Vorliegen von mehr als  $2 \times 2$  Kategorien, also einer mehrdimensionalen  $A \times B$ -Kontingenztafel, wird eine Verallgemeinerung des Phi-Koeffizienten definiert, die als „Cramér's  $V$ “ (oder „ $C$ “) bekannt ist (vgl. Hays, 1977, 745):

$$(9.20) \quad V^2 = \frac{\chi^2}{N (K_{\min} - 1)}$$

( $K_{\min}$  bezeichnet die jeweils geringere Anzahl von Abstufungen oder Modalitäten für die beiden Kategorien  $A$  und  $B$ .)

Auch der Koeffizient  $V^2$  ist im Rahmen des unter (9.18) angegebenen Kontingenzmodells interpretierbar, nicht jedoch als Korrelationskoeffizient (s. o.).

Einen weiteren - wohl den gebräuchlichsten - Ansatz zur Erfassung der Assoziation in Kontingenztafeln stellt der sog. „Kontingenzkoeffizient“  $C$  dar, der wie folgt definiert ist (vgl. Hays, 1977, 745):

$$(9.21) \quad C^2 = \frac{\chi^2}{\chi^2 + N} = \frac{V^2}{1 + V^2}$$

Wie leicht zu erkennen ist, liegt die obere Grenze von  $C^2$  unter Eins, weswegen die Nützlichkeit dieses Maßes im Vergleich zu den anderen häufig angezweifelt wird - vgl. etwa Hays (1977, 745).

Weitere Indices für das Ausmaß an statistischer Assoziation bei nominalen Daten finden sich bei Fleiss (1973), Särndal (1974), Cohen (1977), Glass (1978), Acock & Stavig (1979), Langeheine (1980, 73) sowie Stavig & Acock (1980), ferner in der im Abschnitt 7.5.3.1 genannten Literatur.

## 9.5 Zur Kritik der Maße der statistischen Assoziation

Gegen die in den vorangegangenen Abschnitten dargestellten oder erwähnten normierten Maße der statistischen Assoziation, insbesondere gegen die Korrelationskoeffizienten, sind verschiedentlich Einwände geltend gemacht worden - vgl. dazu insbesondere Glass & Hakstian (1969), Kennedy (1970) sowie Dooling & Danks (1975). Diesen Einwänden sind insbesondere Halderson & Glasnapp (1972), Soderquist & Hussian (1978) sowie Fleishman (1980) entgegengetreten; wir wollen die Argumente und Gegenargumente an dieser Stelle nicht aufgreifen, zumal aus den bisherigen Ausführungen hinreichend deutlich geworden sein sollte, daß das Konzept der experimentellen Effekte - trotz aller möglichen Schwächen - unverzichtbar ist. Drei dieser Schwächen wollen wir im folgenden kurz ansprechen, weil ihre Kenntnis für das Arbeiten mit den EE wesentlich ist.

- (1) Ein gewichtiger Einwand gegen die EE besteht in der Tatsache, daß sie als normierte Größen definiert werden, wobei hierbei entweder Fehler- oder totale Varianzen verwendet werden, deren Größe interexperimentell sehr starken Schwankungen unterworfen sind. Dies bedeutet, daß ein kleiner Effekt als Quotient aus einer großen quadrierten Mittelwertsdifferenz plus einer großen Fehlervarianz entstehen kann oder in einem präziseren Experiment als Quotient aus einem kleinen absoluten Effekt plus einer geringen Fehlervarianz.

Sofern es die *Festlegung* des EEM vor dem Experiment anbelangt, muß diese unter Verwendung eines normierten EE erfolgen, weil die zur Bestimmung des benötigten Stichprobenumfanges (siehe Abschnitt 10) verfügbaren Tabellen der Gütefunktionen aus praktischen Gründen stets un-

ter Verwendung normierter Maße erstellt werden. Dies geschieht analog der Tabellierung der kritischen Werte der Teststatistiken aus den zentralen Verteilungen - man bezieht sich bspw. immer auf die „Standardnormal-Verteilung“ mit  $NV(\mu=0; \sigma_e^2=1)$ . Zudem ist in aller Regel vor dem Experiment nicht bekannt, wie groß  $\sigma_e^2$  sein wird - ohne diese Information können die Tabellen zur Stichprobenumfangsbestimmung nur unter Verwendung normierter EEM benutzt werden - siehe Abschnitt 10.

Darüber hinaus wird sich die Festlegung von absoluten Abweichungen von der Erwartung unter  $H_0$  bei der Planung eines Experiments meist als sehr schwierig erweisen, weil derart genaue Prognosen aus kaum einer wissenschaftlichen Kausalhypothese abgeleitet werden können. Sind dagegen die empirischen Daten erhoben worden, sollte der EES sowohl als Funktion des empirischen Wertes der benutzten Teststatistik wie als absolute Mittelwertsdifferenz oder Treatmentvarianz oder dgl. bestimmt und mitgeteilt werden - dies ist ohne großen Aufwand zu bewerkstelligen, wie die Abschnitte 9.3.2.2 und 9.3.2.3 zeigen.

- (2) Das Ausmaß des EE ist unter sonst gleichen Bedingungen sehr stark abhängig von der Anzahl der untersuchten Treatmentmodalitäten und dem gewählten Versuchsplan. Selbst normierte EE sind interexperimentell nicht direkt miteinander vergleichbar, und festgestellte EES beziehen sich nur auf das durchgeführte Experiment mit seinem speziellen Versuchsdesign und seinen spezifischen Faktormodalitäten. Dennoch ermöglicht eine vergleichende Betrachtung der unter verschiedenen Versuchsbedingungen ermittelten EES eine differenzierte Bewertung der wissenschaftlichen Kausalhypothese, weil etwa festgestellt werden kann, welche Operationalisierungen oder UVn zu relativ kleinen oder aber großen Effekten führen - vgl. Bredenkamp (1980, 52).
- (3) Auf den Einwand, daß die experimentellen Effekte EES - besonders bei kleinen Stichprobenumfängen - mit einem großen Standardfehler behaftet sind, sind wir bereits im Abschnitt 9.3.2.3 eingegangen.

## 9.6 Zusammenfassung

Zusammenfassend und diesen Teil abschließend sei nochmals hervorgehoben, daß das Konzept des experimentellen Effekts - trotz seiner bestehenden Schwächen - als integrativer Bestandteil eines jeden Signifikanztests unverzichtbar ist. Hierfür gibt es vor allem zwei Gründe:

- (1) Die Tatsache der statistischen Signifikanz besagt lediglich, daß der gefundene EE nicht als zufällig zustande gekommen interpretiert werden sollte, nicht jedoch, wie groß er ist. Diese Information ist aber für die folgende Beurteilung der geprüften Kausalhypothese WH wesentlich (vgl. Abschnitt 7.4 und 11.2).



- (2) Bei jedem Signifikanztest kann die Wahrscheinlichkeit  $\beta$  für einen Fehler 2. Art (und damit auch die Teststärke  $1-\beta$ ) nur kontrolliert und/oder bestimmt werden, wenn angegeben werden kann, wie sehr die empirischen Resultate von ihrer Erwartung unter Gültigkeit der  $H_0$  abweichen. Diese Information ist aber von großer Bedeutung, weil bei einem Fehler 2. Art irrtümliche Falsifikationen der WH zu erwarten sind, wenn wir von der Implikationsbeziehung  $WH \rightarrow H_1$  ausgehen.

Zu diesen Punkten haben wir bereits im Zusammenhang mit den Determinanten des Signifikanztests im Abschnitt 7.4.3 ausgeführt, daß der EE als interessierender („praktisch bedeutsamer“) Mindesteffekt EEM zweckmäßigerweise vor dem Experiment vom E nach inhaltlichen Kriterien zu fixieren ist; auf einige Probleme hierbei gehen wir im Abschnitt 11.1 ein. Die Fehlerwahrscheinlichkeit  $\alpha$  und  $\beta$  sind ebenfalls vom E zu kontrollieren und gering zu halten. Trifft der E diese Festlegungen, sollte er den Stichprobenumfang so spezifizieren, daß die Hypothesenprüfung unter Einhaltung der vorgegebenen Kriterien durchgeführt werden kann. Dies gilt natürlich unter der Voraussetzung, daß der tatsächliche EE mindestens so groß ist wie der EEM; andernfalls vermindert sich die Teststärke.

Wie die Stichprobenumfangsbestimmung unter diesen Gesichtspunkten im einzelnen erfolgen kann, werden wir im Teil 10 darstellen.

## 10. Bestimmung des Stichprobenumfanges

### 10.1 Überblick

Unter „Stichprobenumfang“ verstehen wir die Anzahl der Beobachtungs- oder Untersuchungseinheiten ( $V_{pn}$ ; vgl. Abschnitt 3.3.2) pro experimenteller Bedingung (Symbol:  $n$ ) oder pro Experiment (Symbol:  $N$ ).

Mit der Bestimmung von Stichprobenumfängen können grundsätzlich drei verschiedene Ziele verfolgt werden (vgl. Mace, 1964; Enderlein, 1972; Rasch, Herrendörfer & Bock, 1974; Rasch et al., 1978):

1. die Prüfung von statistischen Hypothesen;
2. die (Intervall-)Schätzung für Parameter; und
3. die Lösung von Selektionsproblemen.

Der Bearbeitung von Selektionsproblemen sieht sich der experimentell arbeitende Psychologe u.W. so gut wie nie konfrontiert, weswegen wir diesen Bereich der eher praktischen Arbeit nicht behandeln - eine ausführliche Dar-

stellung findet sich etwa bei Mace (1964, insbes. Kap. 7). Die Prinzipien, die bei der Bestimmung von Konfidenzintervallen für Parameter zu beachten sind, lassen sich relativ einfach aus denjenigen ableiten, die wir im folgenden im Zusammenhang mit der Stichprobengrößenbestimmung zur Prüfung statistischer Hypothesen darlegen werden; wir gehen daher nicht weiter auf die unter 2. genannte Fragestellung ein, zu der Mace (1964), Enderlein (1972) sowie Rasch et al. (1978) detaillierte Informationen geben.

Bei der Bestimmung des „optimalen“ Stichprobenumfanges zur Prüfung von statistischen Hypothesen ist zunächst die Frage zu beantworten, hinsichtlich welchen Kriteriums „Optimalität“ erzielt werden soll. Aus dem sog. „Zentralen Grenzwertsatz“ und dem „Gesetz der großen Zahlen“ (vgl. dazu u.a. Haagen & Pertler, 1976, 158-164; Hays, 1977, 278-283; Pfanzagl, 1978, 61-66) sowie aus unseren Ausführungen zur Präzision einer Untersuchung im Abschnitt 8.4 ergibt sich folgendes Optimalitätskriterium: „Wähle eine möglichst (sehr) große Stichprobe, etwa  $n > 1000!$ “

Stehen dagegen Kostenüberlegungen im Vordergrund, ergibt sich als entgegengesetzte Forderung: „Wähle eine möglichst (sehr) kleine Stichprobe, etwa  $n \leq 5!$ “

Selbstverständlich sind zahlreiche weitere Optimalitätskriterien denkbar, auf die wir jedoch nicht weiter eingehen wollen; man orientiere sich etwa bei Mace (1964) und Rasch, Herrendörfer & Bock (1974).

Aufgrund unserer Erörterungen in den Teilen 6, 7 und 8 ergibt sich die Vorrangigkeit eines anderen Kriteriums: „Wähle die Größe der Stichprobe derart, daß bei vorgegebenem experimentellem Mindesteffekt die Prüfung der statistischen Hypothese bei festgelegten geringen Maximalwerten für die Wahrscheinlichkeiten von Fehlern 1. und 2. Art erfolgen kann!“ Die Erfüllung dieses Kriteriums stellt eine notwendige Bedingung für die strenge Prüfung einer wissenschaftlichen Hypothese dar.

Es wird sich jedoch zeigen, daß sich bei seiner konsequenten Anwendung sehr häufig Stichproben in einer Größenordnung ergeben, die in der Regel nicht zu bewältigen sind - vgl. dazu das erste Beispiel im Abschnitt 10.3.3.2. In diesen Fällen erzwingen allgemeine Kostenüberlegungen i. w. S. (vgl. Pfanzagl, 1978, 106) die Festsetzung einer oberen Grenze für den realisierbaren Stichprobenumfang. Hierdurch ergibt sich oft die Notwendigkeit, die Größe mindestens einer der vier Determinanten des Signifikanztests, die vor der Datenerhebung fixiert wurden, zu modifizieren, und es resultiert möglicherweise eine weniger strenge Prüfung der WH. Ansätze zur Lösung dieses Problems werden teilweise im Abschnitt 10.3.3.2 skizziert und im Abschnitt 11.1.1 nochmals aufgegriffen.

über die unterschiedlichen Möglichkeiten, das Modell der Kosten-Nutzen-Analyse systematisch bei der Bestimmung des Stichprobenumfanges einzusetzen, informieren eingehender die Arbeiten von Cox (1958), Mace (1964, Kap. 11), Overall & Dalal (1965), Cleary & Linn (1969), Dixon & Massey (1969, 89f.), Cochran (1972), Abrahams & Alf (1978) sowie die letzten Jahrgänge des „Biometrical Journal“.

Auf der anderen Seite kann es zuweilen geschehen, daß Stichprobenumfänge in der Größenordnung von  $n = 2$  bis  $n = 4$  berechnet werden. Derartige Zahlen ergeben sich etwa dann, wenn sehr große experimentelle Effekte bei vergleichsweise großen Irrtumswahrscheinlichkeiten aufgedeckt werden sollen. Ist bspw. die zu entdeckende Effektvarianz  $\sigma_b^2$  genauso groß wie die Fehlervarianz  $\sigma_e^2$ , benötigt man für einen mittels t-Test auszuwertenden Zwei-Gruppen-Plan  $n = 4$  Vpn pro Gruppe, wenn die Hypothesenprüfung bei  $\alpha = 0,20$  und  $\beta = 0,20$  zweiseitig erfolgen soll; zur Überprüfung einer gerichteten Hypothese reichen sogar  $n = 3$  Vpn aus. In solchen Fällen muß sich der E die Frage stellen, welche Aussagekraft einem Experiment mit derartig geringen Gruppengrößen angesichts unserer Ausführungen über die Populationsvalidität im Abschnitt 4.1 zukommen kann.

Wir werden in den folgenden Abschnitten die wesentlichsten der bislang entwickelten Ansätze der Stichprobenumfangsbestimmung kurz umreißen, wobei wir das Prinzip ausführlich am Beispiel des t-Tests und des varianzanalytischen F-Tests demonstrieren, während wir für eine Reihe weiterer Testverfahren vornehmlich Hinweise auf die relevante Literatur geben.

Nicht eingehen werden wir dabei auf den nur in recht eingeschränkten Zusammenhängen verwendbaren Ansatz, die Stichprobengrößenbestimmung unter Einbeziehung der Reliabilität der AV vorzunehmen (siehe zu diesem Terminus Abschnitt 2.5 und die dort angegebene Literatur). Die vorgeschlagenen Strategien (vgl. Levin & Subkoviak, 1977, 1978; Subkoviak & Levin, 1977; zur Kritik Forsyth, 1978 a, b) gehen von der Annahme aus, es ließe sich eine eindeutige Beziehung zwischen der Reliabilität eines als AV verwendeten Tests und der Teststärke angeben (vgl. dazu Sutcliffe, 1958; Cleary & Linn, 1969; Cleary, Linn & Walster, 1970; Leeb & Weinberg, 1977). Wie die Befunde von Overall & Woodward (1975, 1976) und Fleiss (1976) sowie die theoretischen Erörterungen von Nicewander & Price (1978) und Bredenkamp (1980) aufzeigen, ist diese Annahme nicht generell gültig (vgl. jedoch Sutcliffe, 1980).

## 10.2 Allgemeine Prinzipien der Stichprobengrößenbestimmung

Zur praktischen Durchführung der Stichprobengrößenbestimmung muß der E vornehmlich Überlegungen zu den folgenden Punkten anstellen:

- (1) Adäquate Umsetzung der wissenschaftlichen Hypothese(n) in statistische
  - siehe Abschnitt 8.1;

- (2) Wahl des ihm unter Berücksichtigung der Teile 1 bis 8 dieser Arbeit geeignet erscheinenden Versuchsplanes mit der Festlegung von Art und Anzahl der Treatmentmodalitäten usw. ;
- (3) Wahl einer Teststatistik - vgl. hierzu die Teile 7 und 8;
- (4) Fixierung der maximalen Wahrscheinlichkeit für einen Fehler 1. Art (Signifikanzniveau  $\alpha$ ) - vgl. Abschnitt 11.12;
- (5) Fixierung der maximalen Wahrscheinlichkeit für einen Fehler 2. Art (oder der Teststärke  $1 - \beta$ ) - vgl. ebenfalls Abschnitt 11.1.2;
- (6) Angabe der Größe des experimentellen Effektes EEM, der mindestens auftreten muß, damit das Resultat als „praktisch bedeutsam“ interpretiert werden kann - siehe Teil 9 und Abschnitt 11.1.3.

Neben diesen Überlegungen und anschließenden Festlegungen benötigt der E die Werte der nicht-zentralen Verteilungen der von ihm gewählten Teststatistik, ohne die die Stichprobenumfangsbestimmung nicht durchführbar ist. Aus diesen nicht-zentralen Verteilungen läßt sich die Gütefunktion des betr. Tests ableiten, die die Abhängigkeitsbeziehungen zwischen den Determinanten dieses Tests verdeutlicht. Diese Abhängigkeitsbeziehungen wiederum lassen sich graphisch oder tabellarisch darstellen (vgl. dazu die o. g. Abschnitte). Betrachten wir beispielhaft nur die Gütefunktion des F-Tests, so sind hierbei im wesentlichen drei verschiedene Typen von Tabellen resp. Graphen anzutreffen.

1. Bei vorgegebenem Signifikanzniveau  $\alpha$  werden die Werte für die Teststärke  $1 - \beta$  in Abhängigkeit von der Größe des experimentellen Effektes (meist als  $\lambda$  oder  $\varphi^2$  angegeben; siehe dazu die folgende Gleichung (10.1)), der Anzahl der Treatment-Modalitäten (i. a. über die Freiheitsgrade  $df_z$  oder  $u$  für den Zähler des F-Bruches aufgelistet) und der Stichprobengröße  $N$  (in der Regel über die Freiheitsgrade  $df_N$  oder  $v$  für den Nenner des F-Bruches angegeben) verzeichnet;
2. es wird die Mindestgröße des experimentellen Effektes aufgeführt, die bei vorgegebenen Werten für die übrigen Determinanten aufgedeckt werden kann; oder
3. es wird der benötigte Mindeststichprobenumfang in Abhängigkeit von den anderen Größen angegeben.

Infolge ihres speziellen Aufbaus eignen sich die Tabellen des unter 3. aufgeführten Typs vornehmlich zur Planung von Untersuchungen, während sich die beiden unter 1. und 2. charakterisierten Typen eher bei der Auswertung von Untersuchungen sinnvoll einsetzen lassen. Hat man ein Experiment aus irgendwelchen Gründen nicht nach dem in dieser Arbeit favorisierten Strategie planen können oder aber ist eine Publikation von experimentellen Ergebnissen so unzulänglich, daß sie keine Angaben über die Determinanten der durchgeführten Signifikanztests enthält, müssen diese Angaben aus den vorhandenen Informationen nachträglich ermittelt werden. Ein zumindest ex post facto

bestimmter Wert für den EES und für die Teststärke dient der adäquaten Interpretation der experimentellen Ergebnisse - siehe dazu insbesondere Abschnitt 11.2.

Wenden wir uns nun dem Umgang mit den genannten Tabellen im einzelnen zu!

### 10.3 Bestimmung des Stichprobenumfanges bei univariaten Varianz- und Regressionsanalysen

Wegen des überwiegenden Gebrauchs, der in der Praxis von den parametrischen Teststatistiken  $t$ ,  $\chi^2$  und  $F$  gemacht wird, sind deren nicht-zentrale Verteilungen am ausführlichsten berechnet, tabelliert und in gekürzter tabellarischer oder graphischer Form publiziert worden, u.a. in Mosteller & Bush (1954), Scheffé (1959), Owen (1962), Pearson & Hartley (1962, 1972), Cochran & Cox (1968), Kirk (1968), Cohen (1969, 1977), Dixon & Massey (1969), Bailey (1971), Winer (1971), Enderlein (1972), Guenther (1973), Rasch, Enderlein & Herrendörfer (1973), Lee (1975), Odeh & Fox (1975) sowie Diehl (1979). Leicht zugänglich sind zudem die ausführlichen Tabellen, die Rotton & Schönemann (1978) veröffentlicht haben.

Um mit diesen Tabellen und Graphen direkt operieren zu können, muß der experimentelle Mindesteffekt EEM in Form eines der Nicht-zentralitätsparameter angegeben werden, die sich auf folgende Definition zurückführen lassen (vgl. auch Abschnitt 9.3.1 und 9.3.2.1):

$$(10.1) \quad \varphi^2 = \frac{\sum_{k=1}^K (\mu_k - \mu)^2 / K}{\sigma_e^2 / n} = \frac{\sigma_b^2}{\sigma_e^2 / n} = n \cdot f^2 = \frac{\lambda}{K}$$

Aus der vorstehenden Aussage folgt jedoch nicht, daß der EEM von vornherein notwendigerweise als Nicht-Zentralitätsparameter festzulegen ist. Die darzustellenden Ansätze der Stichprobenumfangsbestimmung unterscheiden sich im wesentlichen gerade in der Art, in der die erste Festlegung in Abhängigkeit von den verfügbaren Vorinformationen erfolgen kann. Ist bspw.  $\sigma_e^2$  bekannt, muß der EEM nicht als normiertes Maß angegeben werden, wie Formel (10.1) zeigt.

#### 10.3.1 Bei Kenntnis der Populationsvarianz $\sigma_e^2$ („Klassischer Ansatz“)

Bei diesem Bestimmungsverfahren wird davon ausgegangen, daß die EEM als Abweichung der einzelnen Treatment-Mittelwerte  $\mu$  von ihrem Gesamtmittel-

wert  $\mu$  einzeln oder zumindest als Gesamtquadratsumme gemäß der zu prüfenden WH festgelegt werden können. Ist dies geschehen, muß noch eine Angabe über die Größe der Varianz  $\sigma_e^2$  erfolgen.

Wie diese zu erlangen ist, wird in der Literatur entweder nicht angesprochen („... given a preliminary estimate of the standard deviation . . .“; Mace, 1964, 90; „Suppose further that we know that . . . ( $\sigma_e^2$ ) is approximately . . .“; Dixon & Massey, 1969, 278) oder es wird empfohlen,  $\sigma_e^2$  aus Daten zu schätzen, die etwa in Vorversuchen erhoben worden sind (Kirk, 1968, 9; Winne, 1968, 1613; Winer, 1971, 222); vgl. zu diesem Problem auch Feldt (1973).

Hat man sich jedoch auf irgendeine Weise eine Schätzung der Varianz  $\sigma_e^2$  verschaffen können, wählt man die dem vorgegebenem Signifikanzniveau entsprechende Power Chart und setzt nach einer „Versuch-Irrtums-Strategie“ solange verschiedene Werte für  $n$  in Formel (10.1) ein, bis ein Wert für  $\varphi^2$  resultiert, der zur gewünschten Teststärke führt - vgl. zur Vorgehensweise etwa Kirk (1968, 9-11) und Diehl (1979, 26-29). Dieses Verfahren dürfte deshalb in der Praxis nur selten relevant werden, weil erfahrungsgemäß in der Regel Informationen aus gesonderten Vorversuchen oder Pilot-Studien, die unter vergleichbaren Bedingungen durchgeführt wurden, wie dies für den Hauptversuch intendiert ist, nicht verfügbar sind. Beabsichtigt daher der Experimentator, eigene Voruntersuchungen anzustellen, muß er berücksichtigen, daß sich dadurch sein Gesamtversuchspersonenbedarf nicht unwesentlich steigert; er wird vergleichsweise rasch an die Grenzen seiner Möglichkeiten stoßen. Trotz dieser gewichtigen Nachteile wird das soeben vorgestellte Verfahren in der Literatur nicht selten propagiert - vgl. u.a. Kirk (1968, 9f.), Winne (1968), Dixon & Massey (1969, 270-279) sowie Diehl (1979, 25-29).

Das Problem des meist zu großen Stichprobenumfanges ließe sich nicht unwesentlich reduzieren, wenn die in den Vorversuchen erhobenen Daten unverändert auch in den Hauptversuchen Verwendung finden könnten. Ein Verfahren, das dies ermöglicht, wird im folgenden Abschnitt dargestellt.

### 10.3.2 Bei prä-experimenteller Schätzung der Varianz $\sigma_e^2$

(„Two-Stage-Sampling“-Verfahren nach Stein und Rodger)

Auch bei dem „Two-stage-sampling“-Verfahren, das auf Stein (1945) zurückgeht, ist die Populationsfehlervarianz  $\sigma_e^2$  zunächst unbekannt. Um ihre Größe zu ermitteln, wird ein Vorversuch durchgeführt.

Vor diesem wird neben den übrigen obligatorischen Festlegungen der EEM als  $K\sigma_b^2$  angegeben (vgl. dazu Formel (10.1) und Abschnitt 10.2.1); diesen Ausdruck nennt Rodger (1976) „V“.

Anschließend führt der E sein Experiment an  $K$  Stichproben der gleichen, aber willkürlichen und vorläufigen Größe  $n_0$  durch und schätzt aus diesen  $Kn$ , Daten die Varianz  $\sigma_\epsilon^2$ ; diese Schätzung beruht auf den üblichen Freiheitsgraden der Varianzanalyse, nämlich  $df_N = v = K(n_0 - 1)$ . Als nächstes bestimmt der E die Stichprobengröße  $n_1$ , für die gilt (vgl. Rodger, 1976, 3):

$$(10.2) \quad n_1 = \frac{\hat{\sigma}_\epsilon^2 \cdot D\beta}{V} + 1$$

Hierbei kennzeichnet  $D\beta$  einen speziellen Nicht-Zentralitätsparameter, der von Stein (1945) entwickelt wurde und dessen Werte in Abhängigkeit von den Freiheitsgraden  $df_z$  und  $df_N$  erstmalig Rodger (1976) in tabellierter Form vorgelegt hat.

Ergibt die Berechnung von  $n_1$ , daß  $n_1 \leq n_0$ , wird der übliche varianzanalytische F-Test zur Prüfung der generellen Null-Hypothese durchgeführt. Ist jedoch  $n_1 > n_0$ , sind weitere  $n_1 - n_0$  Beobachtungen für jede der  $K$  experimentellen Bedingungen anzustellen.

Zur Bestimmung der Quadratsumme zwischen den experimentellen Bedingungen werden alle  $K \cdot n_0$  Beobachtungen verwendet; die statistische Prüfung erfolgt allerdings gegen die eingangs bestimmte Varianz  $\hat{\sigma}_\epsilon^2$ . Der EES wird nach der Datenerhebung aus den Daten bestimmt.

Die Stichprobengröße  $n_1$  hängt bei diesem Verfahren u.a. von der Wahl des Wertes für  $n_0$  ab; Rodger (1967, 192) bemerkt, daß ein Nachteil des zweistufigen Stichprobenverfahrens darin liegt, daß der Wert für  $n_1$  nach der ersten Stufe sich als „unpraktikabel groß“ herausstellen kann. Diese Schwierigkeit kann zuweilen durch die Wahl eines „optimalen“ Wertes für  $n_0$  umgangen werden; Hinweise hierzu entnehme man Rodger (1967, 192) und Bishop (1978).

Beispiele zur praktischen Durchführung dieses Verfahrens sowie weitere Einzelheiten finden sich in Rodger (1967, 1976), Guenther (1973, 217-220) und Bishop (1978); zu einem ganz ähnlichen Ansatz siehe Keppel (1973, 536-540).

Hinweise zur Bestimmung des Stichprobenumfanges für multiple Vergleiche nach einer (signifikanten) Varianzanalyse unter Verwendung der „Two-stage-sampling“-Prozedur geben die Arbeiten von Rodger (1974, 1975a, b, 1978) und von Hochberg & Lachenbruch (1976).

### 10.3.3 Ohne Kenntnis der Populationsvarianz $\sigma_e^2$

Bei den im folgenden zu skizzierenden Verfahren wird das Problem, vor dem Experiment die Größe von  $\sigma_e^2$  nicht zu kennen, dadurch umgangen, daß der E den EEM als relativen Anteil der Effekt- an der Fehler- oder totalen Varianz angibt. Diese Art der Festlegung ist insofern als unspezifisch zu bezeichnen, als nicht mehr einzelne Mittelwertsdifferenzen oder entsprechende Quadratsummen als „praktisch bedeutsam“ angesetzt werden können, sondern nur noch eine (globale) relative „mittlere Differenz“. Andererseits dürfte eine derartige Festlegung noch am ehesten den in der Regel ebenfalls recht unspezifischen psychologischen Hypothesen entsprechen - vgl. Abschnitt 8.1.1.

#### 10.3.3.1 Festlegung von $\varphi^2$

Je nach persönlichen Präferenzen trifft der E bei der hier anzusprechenden Vorgehensweise eine globale Festlegung für die Mindestgröße von  $\varphi^2$  (vgl. Formel (10.1)) oder aber für  $f^2$  resp.  $R_{Y.X}^2$  (siehe dazu Formel (9.4) und (9.9)); die letztgenannten Indices sind anschließend in  $\varphi^2$  umzurechnen (vgl. Formel (9.10) und (10.1)).

Unter Verwendung der entsprechenden Teststärke-Tabellen kann man dann die benötigte Stichprobengröße nach einem „Versuch-Irrtum-Verfahren“ ermitteln - vgl. zur Durchführung etwa Diehl (1979, 26-36) - oder aber direkt ablesen, wenn man günstiger aufgebaute Tabellen benutzt, wie sie etwa in Odeh & Fox (1975) zu finden sind.

Beispiele für diese und einige aus ihr abgeleitete Bestimmungsstrategien findet man u.a. in den Publikationen von Kirk (1968, 109f.), Levin (1972, 1975), Guenther (1973, 209f., 355-359), Keppel (1973, 529-541) und Diehl (1979, e. g. 29-39).

Es empfiehlt sich bei diesem und den vorhergehenden Ansätzen grundsätzlich, die Tabellen den Graphen vorzuziehen, weil bei letzteren die Ablesegenauigkeit geringer ist.

#### 10.3.3.2 Festlegung von $f^2$ oder $R_{Y.X}^2$ (Verfahren nach Cohen)

Bei dem abschließend anzusprechenden Verfahren handelt es sich um den Teststärkenanalyse-Ansatz, der von Cohen (1962, 1965, 1969, 1970, 1973b, 1977; Cohen & Cohen, 1975) entwickelt worden ist.

Cohen kommt im Vergleich zu zahlreichen anderen Autoren (s. o.) mit einem leicht zugänglichen Minimum an Tabellen aus, die zudem so angelegt sind, daß



die Stichprobenumfangsbestimmung sowohl für einfache und komplexe varianzanalytisch wie für praktisch alle regressionsanalytisch auszuwertenden Versuchspläne relativ einfach vorgenommen werden kann.

Für die varianzanalytische Prüfung von Hypothesen über Mittelwerte verwendet Cohen das bereits im Abschnitt 9.3.1 eingeführte  $f^2$  (vgl. auch Formel (10.1)) als Index für den experimentellen Effekt; für die entsprechende regressionsanalytische Prüfung von Hypothesen über multiple Korrelationsquadrate wird das Maß  $R^2_{Y.X}$  benutzt - siehe dazu Abschnitt 9.3.2.1.

Neben der Verwendung bei der Planung von Experimenten können die in Cohen (1969, 1977) publizierten Tabellen auch zur Auswertung von Experimenten, insbesondere zur nachträglichen Bestimmung der Teststärke, herangezogen werden. Dies ist insofern vorteilhaft, als die dazu verfügbaren Computer-Programme (vgl. etwa Woodward & Overall, 1976) nur recht begrenzt verwendbar sind.

Wir wollen im folgenden trotz des begrenzten uns zur Verfügung stehenden Raumes den Gebrauch dieser Tabellen Cohens an einem Beispiel aufzeigen. Die Beschäftigung mit diesem Beispiel kann selbstverständlich die Lektüre des Buches von Cohen (1977) nicht ersetzen; sie kann jedoch deutlich machen, wie problemlos sich die Stichprobenumfangsbestimmung vornehmen läßt.

Gehen wir der Einfachheit halber davon aus, daß ein Zwei-Gruppen-Experiment zur Prüfung der folgenden statistischen Hypothese  $H_0$  geplant ist:

$$H_0: \mu_1 - \mu_2 \leq 0 \text{ gegen die } H_1: \mu_1 - \mu_2 > 0.$$

Die notwendigen Voraussetzungen (siehe Abschnitt 8.2) werden als erfüllt vorausgesetzt, so daß der t-Test zur Anwendung kommen kann. Der E hat die folgenden Festlegungen getroffen:

$$\alpha = 0,01; \beta = 0,01; R^2_{Y.X} = 0,0099 \text{ (entspricht } f^2 = 0,01).$$

Es sollen die beiden Hypothesen bei gleichen maximalen Fehlerwahrscheinlichkeiten „fair“ gegeneinander getestet werden (vgl. dazu Abschnitt 11.1.2); die bei Fehlentscheidungen möglichen Folgen werden als gleich schwerwiegend angesehen. Es wird eine quadrierte und normierte Mittelwertsdifferenz dann als praktisch bedeutsam interpretiert, wenn sie mindestens 0,99% der Gesamt- oder 1% der Fehlervarianz ausmacht. Diese Werte müssen für die Benutzung der Tabelle 2.4.1 (Cohen, 1977, 54f.) zunächst in den von Cohen gewählten Index  $d$  umgerechnet werden, der wie folgt definiert ist:

$$(10.3) \quad d = \frac{\mu_1 - \mu_2}{\sigma_e} = 2f$$

Aus  $f = 0,1$  folgt, daß  $d = 0,2$ ; es handelt sich nach Cohens vorgeschlagener Konvention (1977, 24-27) um einen „kleinen Effekt“.

Die vorgegebenen Informationen reichen aus, um direkt aus dem mit „ $\alpha_1 = 0,01$ “ überschriebenen Teil der gen. Tabelle den pro experimenteller Bedingung benötigten Stichprobenumfang abzulesen:

$$n = 1084 \text{ und somit insgesamt } 2n = N = 2168 \text{ Vpn.}$$

Die bei der Größe  $\alpha$  vorgenommene Indizierung mit „1“ besagt, daß ein einseitiger Test zugrunde gelegt wird; bei einem vorgesehenen zweiseitigen Test (einer ungerichteten Hypothese) hätte die Ablesung des Wertes für  $n$  unter dem mit „ $\alpha_2 = 0,01$ “ überschriebenen Teil der Tab. 2.4.1 vorgenommen werden müssen. Es gilt hier wie auch für die Varianz- und regressionsanalytischen Tabellen, daß einem einseitigen Test bei  $\alpha_1 = c$  ein zweiseitiger Test bei  $\alpha_2 = 2c$  entspricht und umgekehrt.

üblicherweise ist ein E nicht in der Lage, die soeben bestimmte Anzahl von Vpn verfügbar zu machen. Wenn wir davon ausgehen, daß eine Vergrößerung des zu entdeckenden praktisch bedeutsamen Mindesteffektes zunächst nicht in Frage kommt, kann sich der E mit der Erhöhung der maximalen Irrtumswahrscheinlichkeiten behelfen (vgl. zu dieser Empfehlung jedoch auch Abschnitt 8.3). Er kann also etwa festlegen:

$$\alpha = 0,10 \text{ und } \beta = 0,20.$$

Diese Festlegung impliziert, daß die Risiken, die mit einer Fehlentscheidung zugunsten der statistischen Null-Hypothese verbunden sind, als halb so schwerwiegend angesehen werden wie die, die bei einer falschen Entscheidung für  $H_1$  entstehen können - vgl. Abschnitt 11. Dem mit „ $\alpha_1 = .10$ “ überschriebenen Teil der Tab. 2.4.1 ist unter „d = .20“ und „Power = .80“ zu entnehmen, daß  $n = 226$ ; also sind insgesamt 452 Vpn den experimentellen Behandlungen zu unterziehen. Sind auch diese Versuchspersonenmengen für den Experimentator nicht verfügbar, kann er sich in dieser Situation nur noch auf die Entdeckung eines größeren experimentellen Effektes festlegen, wenn er die Fehlerwahrscheinlichkeiten nicht weiter vergrößern will und wenn eine andere Prüfung der wissenschaftlichen Hypothese nicht möglich ist.

Nehmen wir an, der E entschließt sich, einen Effekt dann als psychologisch bedeutsam zu interpretieren, wenn die Effektvarianz (zwischen den Mittelwerten) mindestens 16% der Fehlervarianz oder 13,79% der Gesamtvarianz beträgt, es sich also nach Cohens vorgeschlagener Konvention um einen „großen Effekt“ handelt (Cohen, 1977, 24-27). Aus dieser Festlegung ergeben sich die folgenden Werte:

$$f^2 = 0,16; R^2_{Y.X} = 0,1379; d = 0,8$$

Unter Verwendung der o.a. liberalen Kriterien bzgl.  $\alpha$  und  $\beta$  läßt sich der entsprechenden Tabelle entnehmen, daß der E jetzt nur noch  $n = 14$  oder insgesamt 28 Vpn benötigt.

Für die zweiseitige Prüfung einer ungerichteten Hypothese müßten unter den gleichen Bedingungen  $n = 20$  ( $N = 40$ ) Vpn an dem Experiment teilnehmen.

In ähnlicher Weise, wie es hier am Beispiel aufgezeigt wurde, gestaltet sich auch die Bestimmung des Stichprobenumfanges für einen mittels t-Test auszuwertenden Versuchsplan mit  $n$  Paaren von abhängigen Messungen - siehe dazu im einzelnen Cohen (1977, Kap. 2).

Für Varianz- oder regressionsanalytisch auszuwertende ein- oder mehrfaktorielle Versuchspläne finden sich entsprechende Tabellen auf den Seiten 440-442 des Buches von Cohen (1977); ihre Benutzung gestaltet sich am einfachsten, wenn der experimentelle Effekt als partielles oder semi-partielles multiples Korrelationsquadrat  $R^2_{Y.X}$  (vgl. dazu insbesondere Kerlinger & Pedhazur, 1973, Kap. 5; Cohen & Cohen, 1975, e.g. 78-84, 129-134) festgelegt werden kann. Der nachträglichen Bestimmung der Teststärke aus bereits erhobenen Daten dienen die analog aufgebauten Tabellen auf den Seiten 416-418 (Cohen, 1977).

Die Tabellen in Cohen (1977, Kap. 9) eignen sich zur Bestimmung der benötigten Stichprobenumfänge nicht nur für varianzanalytische Versuchspläne, sondern darüber hinaus auch für Trendanalysen (vgl. Cohen & Cohen, 1975, Kap. 6; Bredenkamp, 1980, 56-59), Meßwiederholungspläne (Cohen & Cohen, 1975, Kap. 10; Bredenkamp, 1980, 91-99), Parallelisierungspläne (Cohen & Cohen, 1975, Kap. 10; Bredenkamp, 1980, 71-73) und Kovarianzanalysen (Cohen, 1977, 379f.; Bredenkamp, 1980, 78-80).

## 10.4 Hinweise zur Stichprobenumfangsbestimmung bei weiteren Gruppen von parametrischen Testverfahren

### 10.4.1 Varianzanalyse mit zufälligen und gemischten Effekten

Die Unterscheidung zwischen drei varianzanalytischen Modellen, die wir kurz im Abschnitt 8.2.1 angesprochen haben, ist sowohl für die Auswertung als auch für die Planung des Experiments von Bedeutung.

Zuweilen bezieht sich eine wissenschaftliche Hypothese auf eine (sehr) große Menge möglicher Treatmentmodalitäten, die der E aus verschiedenen Gründen nicht in ihrer Gesamtheit untersuchen kann. In diesem Fall besteht für den E die Möglichkeit, aus dieser Population von Bedingungen eine Zufallsstichprobe zu ziehen oder aber die realisierten Modalitäten als eine solche zu interpretieren. Nach der Auffassung zahlreicher Autoren (e. g. Bredenkamp, 1975, 810; Hays, 1977, 542) rechtfertigt dieses Vorgehen die statistische Verallgemeinerung der Resultate auch auf alle nicht untersuchten Bedingungen, die in der WH enthalten sind.

Ein derartiger Faktor ist im Sinne der varianzanalytischen Modelle „zufällig“ (vgl. Abschnitt 8.2.1); üblicherweise stellt der Versuchspersonen-Faktor bei Meßwiederholungen einen zufälligen Faktor dar (vgl. Abschnitt 8.4.3).

Wir haben im Abschnitt 2.3 Gründe angegeben, die in aller Regel eine systematische Auswahl von Modalitäten nahelegen. Darüber hinaus ist der F-Test beim Modell der zufälligen Effekte, bei dem Hypothesen über Varianzen, nicht über Mittelwerte geprüft werden, sehr empfindlich gegenüber Verletzungen seiner Voraussetzungen, insbesondere der der Normalverteilung der

Modalitäten des zufälligen Faktors in der Population (vgl. Lehmann, 1968, 45; Glass & Stanley, 1970, 462; Hays, 1977, 540f., sowie Abschnitt 8.2.4). Aus diesen Gründen halten wir insgesamt das zufällige und das gemischte Modell der VA (vgl. Abschnitt 8.2.1) für wenig empfehlenswert, sofern seine Verwendung nicht aufgrund der Fragestellung zwingend indiziert ist; bspw. wird der Versuchspersonen-Faktor bei Meßwiederholungen univariat varianzanalytisch i.a. unter Annahme dieses Modells ausgewertet.

Entschließt sich der E, trotz der angeführten Bedenken dem Modell der zufälligen Effekte den Vorzug zu geben, muß er die Auswertung der Daten entsprechend vornehmen, da teilweise andere Prüfvarianzen im Nenner des F-Bruches resultieren als beim Modell der fixierten Effekte - vgl. zu den Einzelheiten etwa Hays (1977, Kap. 13).

Darüber hinaus folgt der F-Bruch bei zufälligen Effekten auch unter Gültigkeit der statistischen Alternativhypothese den *zentralen* F-Verteilungen. Aufgrund dieser Tatsache muß auch die Bestimmung der Teststärke und der benötigten Stichprobengröße nicht mehr über die nicht-zentralen, sondern über die zentralen F-Verteilungen erfolgen. Auf die Einzelheiten gehen wir nicht ein, sondern verweisen auf John (1971, 53f.), Lindman (1975, 118-135), Hays (1977, 539f.) und Henning & Muthig (1979, 241-248), die das Vorgehen demonstrieren.

#### 10.4.2 Nicht-orthogonale Varianzanalysen

Die bisherigen Stichprobengrößenbestimmungen erfolgten stets derart, daß gleiche Umfänge pro Treatmentmodalität (und pro Zelle des Versuchsplanes) resultierten. Da bei psychologischen Experimenten in der Regel davon ausgegangen werden kann, daß der Erwartungswert der Korrelation zwischen den untersuchten Faktoren oder UV, gleich Null ist, stellt die Gleichbesetzung der Zellen mit Untersuchungseinheiten das adäquate Vorgehen dar (vgl. etwa Yates, 1933).

Bei der Datenerhebung kann jedoch aus den verschiedensten Gründen der Fall eintreten, daß Vpn ausfallen („experimentelle Mortalität“), so daß ungleiche Zellenbesetzungen resultieren. Dieser Ausfall hat zur Folge, daß die Korrelation zwischen den experimentellen Faktoren nicht mehr gleich Null ist (vgl. zur genaueren Charakterisierung der Nicht-Orthogonalität etwa Steyer, 1979, e.g. 82). In diesem Fall sind auch die Quadratsummen der Varianzanalyse nicht mehr additiv (vgl. Steyer, 1979, 79f., 89f.).

Die Frage, auf welche Art Experimente auszuwerten sind, deren Faktoren nicht orthogonal zueinander sind, wird - ausgehend von einem Artikel von Overall & Spiegel (1969) - kontrovers diskutiert. Die wesentlichsten Aspekte dieser Diskussion hat Steyer (1979) zusammengefaßt und vertieft (vgl. auch Bredenkamp, 1980, 74-78).

Unter den Publikationen neueren Datums sind insbesondere Milliken & McDonald (1976), Hamer & Hosking (1977), Lewis & Keren (1977), Blair & Higgins (1978), Herr

& Gaebelein (1978), Jennings (1978) sowie Cramer & Appelbaum (1980) zu erwähnen; empirische Untersuchungen zur Angemessenheit der verschiedenen in der Literatur vorgeschlagenen Auswertetechniken haben u.a. Levy, Narula & Abrami (1975), Narula, Abrami & Levy (1976) (zur Kritik dieser Arbeiten siehe Hummel, 1977) und Steyer (1979) vorgelegt.

Den genannten Arbeiten ist zu entnehmen, daß die Verwendung der üblichen approximativen varianzanalytischen Techniken (vgl. u.a. Winer, 1971) sehr häufig nicht anzuraten ist. Statt dessen sollte auf die regressionsanalytischen Verfahren zurückgegriffen werden, die die Korrelationen zwischen den UVn zu erfassen erlauben und in der Regel exakte Tests der geprüften Hypothesen ermöglichen. Die Korrelationen können dabei je nach Ausgangslage entweder auspartialisiert oder aber in die Auswertung einbezogen werden; bzgl. der Einzelheiten verweisen wir insbesondere auf Bredenkamp (1975, 809; 1980, 74-79), ferner auf die entsprechenden Kapitel in Gaensslen & Schubö (1973, 1976), Kerlinger & Pedhazur (1973) sowie Cohen & Cohen (1975).

Für die Interpretation der Daten ist ungeachtet der statistischen Auswertung jedoch zu bedenken, daß ein nicht-zufälliger (d.h. z.B. von der Art des Treatments abhängiger) *Ausfall* von Vpn zu Beeinträchtigungen der internen Validität führt (vgl. Jurs & Glass, 1971; sowie Abschnitt 3.3.2), während eine *nachträgliche Orthogonalisierung* der Faktoren die Populationsvalidität herabsetzen kann (siehe Jurs & Glass, 1971; Bredenkamp, 1975, 809; und Abschnitt 4.1).

#### 10.4.3 Multivariate Varianz- und Regressionsanalysen

Werden in einem Experiment die Auswirkungen der unabhängigen Variable(n) nicht nur auf eine, sondern auf mehrere abhängige Variable untersucht, liegt eine multivariate Versuchsanordnung vor (vgl. Abschnitt 1.1). Wir haben im Abschnitt 8.4.6.2.2 einige Gründe genannt, die zur Auswertung von multivariaten Versuchsplänen das Testkriterium V von Bartlett (1937, 1939) und Pillai (1955) nahelegen.

Dieses V stellt die Summe der kanonischen Korrelationsquadrate dar; sind die Prädiktorvariablen orthogonal, ergibt sich V als die Summe der multiplen Korrelationsquadrate zwischen je einer der UVn  $X_z$  und jeweils allen AVn  $Y_q$  (vgl. Kerlinger & Pedhazur, 1973; Bredenkamp, 1980).

Für die Stichprobenumfangbestimmung für Tests von multiplen Korrelationsquadraten können die entsprechenden Tabellen von Cohen (1977, 440-442) verwendet werden, so daß sich unter Verwendung des V-Kriteriums keine grundsätzlichen Schwierigkeiten bei der Planung multivariater Versuchspläne ergeben. Welche Überlegungen hierbei im einzelnen anzustellen sind und wie zweckmäßigerweise vorzugehen ist, schildert ausführlicher Bredenkamp (1980).

Bei der Planung des Stichprobenumfangs für ein einfaktoriell-multivariates Experiment, das mittels des  $T^2$ -Tests nach Hotelling (1931) ausgewertet werden soll, können zudem die von Läuter (1978) vorgelegten Tabellen verwendet werden.

Entschließt man sich zur Anwendung eines der übrigen multivariaten Analyseverfahren (vgl. dazu die im Abschnitt 7.5.3.3 angegebene Literatur), kann die Stichprobenumfangsbestimmung derzeit nur approximativ über die o.a. Tabellen erfolgen, weil ausführliche exakte Tabellen für diese Verfahren noch nicht verfügbar sind (vgl. Olson, 1974; Stevens, 1980).

## 10.5 Hinweise zur Stichprobenumfangsbestimmung bei nicht-parametrischen Verfahren

### 10.5.1 Nominale Daten

Zur Auswertung nominaler Daten, die etwa in Kontingenztafeln angeordnet sind, wird in Abhängigkeit von der Anzahl der untersuchten Merkmalskategorien vorwiegend die Binomial- oder die Multinomialverteilung verwendet - vgl. Abschnitt 7.5.3.1.

Ist die Anzahl der zu untersuchenden Merkmale so festgelegt, daß zur Auswertung einer der Binomialtests in Frage kommt, kann die Bestimmung des (insgesamt) benötigten Stichprobenumfanges über die speziellen Tabellen erfolgen, die etwa Fleiss (1973, 176-194, ferner Kap. 3), Cohen (1977, Kap. 4 und 5) sowie Bortz, Österreich & Vogelbusch (1979) vorgelegt haben. Auch für einen Vorzeichen-Test läßt sich der Stichprobenumfang unter Verwendung dieser Tabellen bestimmen.

Will man die entsprechende Auswertung approximativ über einen der vielseitigen  $\chi^2$ -Tests vornehmen, die sowohl zur Analyse der o. gen. Kontingenztafeln als auch u.a. zur Prüfung der Anpassungsgüte verwendet werden können, findet man bei Cohen (1977, Kap. 7) ebenfalls entsprechende Tabellen zur Ermittlung des benötigten Mindeststichprobenumfanges.

Die Bestimmung des im Experiment aufgedeckten EE erfolgt jeweils unter Verwendung eines der im Abschnitt 9.4.2 vorgestellten resp. angesprochenen Maße.

### 10.5.2 Ordinale Daten

Für die insgesamt recht zahlreichen Techniken zur Auswertung von ordinalen oder Rangdaten liegen Tabellen zur Bestimmung des Stichprobenumfanges

und damit allgemein zur Teststärkenanalyse derzeit noch nicht vor, weil die Ableitung der zu ihrer Erstellung benötigten nicht-zentralen Verteilungen von Rangstatistiken noch weitgehend aussteht (vgl. jedoch zu ersten Ansätzen Kraemer, 1974 und Henze, 1979).

Dennoch sollte man auch bei der vorgesehenen Verwendung dieser Gruppe von Verfahren nicht auf die notwendigen Festlegungen bzgl. der Maximalwerte für beide Fehlerarten und den experimentellen Effekt verzichten.

Das Festlegen des experimentellen Mindesteffektes und die Determination des Stichprobenumfanges kann dann näherungsweise über die für das analoge parametrische Verfahren benutzten Effektgrößenindices und Bestimmungstabellen vorgenommen werden.

Diese vorläufige Lösung scheint uns gerechtfertigt, wenn man sich vergegenwärtigt, daß nicht-parametrische Tests fast stets dann effizienter sind als ihre parametrischen Homologa, wenn die zur Anwendung der letzteren notwendigen Voraussetzungen nicht gegeben sind (vgl. zu Einzelheiten die Abschnitte 7.5.3.2 und 8.2).

Will man mit großer Sicherheit eine festgesetzte Mindestteststärke nicht unterschreiten, empfiehlt sich folgendes Vorgehen: Man bestimmt den bei vorgegebenem Signifikanzniveau, Mindesteffekt und festgelegter Teststärke benötigten Stichprobenumfang über die entsprechenden parametrischen nicht-zentralen Verteilungen, wie sie etwa in Cohen (1977) tabelliert sind. Das Ergebnis dividiert man durch die für das vorgesehene nicht-parametrische Verfahren angegebene Maßzahl der relativen asymptotischen Effizienz, die der im Abschnitt 7.5.3 angegebenen Literatur entnommen werden kann - Beispiele für diese Vorgehensweise finden sich in Bredenkamp (1980, 61 f., 73f.). Der experimentelle Effekt läßt sich anschließend etwa unter Verwendung eines der im Abschnitt 9.4.1 angegebenen Maße abschätzen.

## 10.6 Abschließende Bemerkungen zur Stichprobengrößenbestimmung

Der vorangegangene Teil 10 dieses Artikels sollte den Planer und Auswerter von psychologischen Experimenten mit den unterschiedlichen Ansätzen bekannt machen, aufgrund derer die Größe von Stichproben nach einem vorgegebenen Kriterium bestimmt werden kann.

Den vorgestellten Verfahren ist gemeinsam, daß sie auf die Kontrolle und Minimierung der Wahrscheinlichkeiten für falsche Entscheidungen abzielen. Man kann derartige Ansätze als einen wesentlichen Aspekt des umfassenderen

Konzepts der Teststärkenanalyse auffassen, über die Cohen (1977, Kap. 1) ausführlich informiert.

Zu den vielseitigen Möglichkeiten zur praktischen Anwendung dieses Konzepts zählen dabei nicht nur die hier skizzierten Verfahren zur Ermittlung von Stichprobenumfängen, sondern auch die mittels der Tabellen Cohens (1977) leicht mögliche Bestimmung der Teststärke bei vorgegebener Stichprobengröße, also nach der Datenerhebung und statistischen Hypothesenprüfung. Diese Anwendung ist als wesentlicher Bestandteil von statistischen Reanalysen anzusehen.

Abschließend bleibt nur noch zu hoffen, daß Verfahren der Art, wie sie hier vorgestellt worden sind, in absehbarer Zeit die sporadisch propagierten „Faustregeln“ (vgl. etwa Cowles, 1974; Pfanzagl, 1978, 101-106; McGuigan, 1979, 287) ersetzen werden.

## *11. Eine Strategie zur Entscheidung über wissenschaftliche Hypothesen mittels Signifikanztests*

### 11.1 Stadium der Planung des Experiments

#### *11.1.1 Überblick*

Wir gehen im folgenden davon aus, daß der Experimentator (E) aus der wissenschaftlichen Hypothese WH eine statistische Alternativhypothese  $H_1$  abgeleitet hat, unsere Überlegungen sind aber auch auf jene selteneren Fälle zu übertragen, in denen eine Nullhypothese impliziert wird (vgl. Teil 6 und 8.1). Soll eine statistische Hypothese über einen Signifikanztest geprüft werden, hat sich der E auch mit den Determinanten eines solchen Tests auseinanderzusetzen: das Signifikanzniveau  $\alpha$ , die Wahrscheinlichkeit  $\beta$  für einen Fehler 2. Art, den tatsächlichen experimentellen Effekt EEP sowie die Präzision des Experiments, die definiert ist durch den Standardfehler der benutzten Teststatistik, der seinerseits von der Stichprobengröße N und (bei parametrischen Verfahren) von der Fehlervarianz  $\sigma_c^2$  abhängt. Zwischen diesen Größen bestehen eine Reihe von Beziehungen, die im Teil 7 dargestellt worden sind. Dort ist auch bereits deutlich geworden, daß viele empirisch arbeitende Psychologen diesen Abhängigkeitsbeziehungen nur geringe oder gar keine Beachtung schenken. Insbesondere führt die Mißachtung der Fehlerwahrscheinlichkeit  $\beta$  dazu, daß Entscheidungen über statistische Hypothesen getroffen werden, die mit einer relativ hohen Wahrscheinlichkeit falsch sind. Da dies sowohl zu falschen Schlüssen über die Gültigkeit psychologischer Hypothesen und Theorien führen kann wie zu ungerechtfertigten Handlungsvorschlägen für praktische Si-



tuationen, ist die fortlaufende Mißachtung fundamentaler Gesetzmäßigkeiten bei der Anwendung von Signifikanztests nicht zu tolerieren. Dies um so weniger, als bereits mit geringem Aufwand dafür Sorge getragen werden kann, daß der relative Anteil von falschen Entscheidungen aufgrund von statistischen Hypothesenprüfungen vom Forscher selbst durch die Kontrolle *beider* statistischen Fehlerwahrscheinlichkeiten niedrig gehalten werden kann. Dieses Ziel wird durch folgende *Planungsstrategie* erreicht:

Der Experimentator legt die ihm als maximal zulässig erscheinende Irrtumswahrscheinlichkeit 1. Art,  $\alpha$ , auf einen kleinen Wert fest, um auf diese Weise die Gefahr gering zu halten, ein fälschlich die Hypothese WH bestätigendes Ergebnis zu erhalten.

Er fixiert die Wahrscheinlichkeit  $\beta$  für einen Fehler 2. Art ebenfalls auf einen kleinen Wert, um die Gefahr fälschlicher Falsifikationen der WH gering zu halten.

Er gibt an, wie groß der experimentelle Effekt EEM mindestens sein muß, um für die Prüfung der WH relevant („praktisch bedeutsam“) zu sein (s. Teil 9).

Der Experimentator kann dann den Stichprobenumfang  $N$  bestimmen, der notwendig ist, um den erwünschten Mindesteffekt EEM bei festgelegter Teststärke  $1 - \beta$  und gewähltem Signifikanzniveau  $\alpha$  als signifikant ausweisen zu können, falls dieser tatsächlich vorliegt (s. Teil 10).

Hält der Experimentator die ermittelte notwendige Stichprobengröße  $N$  aus ökonomischen Gründen (im weitesten Sinne) für zu groß, sollte er die Werte für  $\alpha$ ,  $\beta$  und EEM unabhängig voneinander und in verschiedenen Kombinationen erhöhen und dazu das jeweils notwendige  $N$  bestimmen. Nach den Erfordernissen des Einzelfalles kann er sich dann für eine Kombination von vertretbaren Werten für diese vier Größen entscheiden. Diese Entscheidung sollte samt ihrer Begründung in jedem Forschungsbericht explizit erwähnt werden.

In einigen Ausnahmefällen dagegen kann die vorgestellte Strategie zu dem Ergebnis führen, daß weniger als etwa fünf  $V_{pn}$  in jeder experimentellen Bedingung ausreichend sind (vgl. Abschnitt 10.1). Wir halten es bei derart kleinen Stichprobenumfängen für kaum möglich, zu einer hinreichenden Sicherung der internen Validität (vgl. Abschnitt 3.3.1) und der Populationsvalidität (vgl. Teil 4) zu gelangen.

### 11.1.2 Zur Festlegung der beiden Fehlerwahrscheinlichkeiten

Bezüglich der Festlegung der Wahrscheinlichkeit für einen Fehler 1. Art haben sich bestimmte Konventionen durchgesetzt, nämlich die überwiegende Ver-

Wendung von  $\alpha = 0,05$  und  $\alpha = 0,01$ . Diese Präferenz ist z.T. darauf zurückzuführen, daß die kritischen Werte auch der häufiger gebrauchten Teststatistiken oft nur für diese Wahrscheinlichkeiten veröffentlicht werden. Zahlreiche Simulationsstudien zur Robustheit parametrischer Tests (vgl. Abschn. 8.2) legen jedoch nahe, das Signifikanzniveau *nicht ohne zwingenden Grund* zu klein werden zu lassen, falls die Vermutung besteht, daß erhebliche Abweichungen von der Normalverteilungsannahme vorliegen. Die Stichprobenverteilungen der meisten parametrischen Teststatistiken sind nämlich besonders in den extremen Wertebereichen („Enden“) sehr empfindlich gegenüber derartigen Abweichungen (vgl. Abschn. 7.5.1). Man sollte deshalb eher  $\alpha = 0,05$  als  $\alpha = 0,01$  wählen; dies gilt insbesondere bei einseitigen Tests. Eine Erhöhung des Signifikanzniveaus etwa auf  $\alpha = 0,10$  oder  $\alpha = 0,20$  kann notwendig werden, wenn der berechnete Stichprobenumfang praktisch nicht zu realisieren ist (s. o.). Dagegen kann eine Verringerung des  $\alpha$ -Niveaus angezeigt sein, wenn zur Prüfung einer wissenschaftlichen Hypothese mehrere statistische Hypothesenprüfungen durchgeführt werden (s. Abschn. 8.3).

Für die Festlegung der maximalen Wahrscheinlichkeit für einen Fehler 2. Art gibt es keine vergleichbar weithin anerkannten Konventionen. Cohen (1965; 1969, 51-54) hat jedoch vorgeschlagen, die Teststärke  $1 - \beta$  nicht geringer als 0,80 werden zu lassen. Dies bedeutet, daß „auf lange Sicht“ stets vier von fünf zutreffenden Alternativhypothesen aufgrund der experimentellen Daten als richtig ausgewiesen werden sollten (vgl. auch Brewer, 1972, 394). Einem „fairen“ Test entspräche die Festlegung  $\alpha = \beta$  (vgl. Bredenkamp, 1969b, 283; Bailey, 1971, 333-336). Hierbei haben beide statistischen Hypothesen unter ihrer Richtigkeit die gleiche Chance, sich zu bewähren. Wählt man dabei allerdings beide Wahrscheinlichkeiten sehr klein (etwa  $\alpha = \beta = 0,01$ ), hat diese Festlegung in der Regel zur Folge, daß sehr große Stichproben untersucht werden müßten (vgl. Abschn. 10.3.3.2). Bei der daher meist unumgänglichen Festlegung unterschiedlicher Werte für  $\alpha$  und  $\beta$  muß bedacht werden, daß damit zumindest implizit eine Gewichtung der zu prüfenden Hypothesen vorgenommen wird. Legen wir bspw. die Werte  $\alpha = 0,05$  und  $\beta = 0,20$  fest, wird damit ausgesagt, daß ein Fehler 1. Art viermal so schwer wiegt wie ein Fehler 2. Art (vgl. hierzu im einzelnen Cohen, 1977, 56; Krause & Metzler, 1978, 254f.; ferner Lehmann, 1958 sowie Rasch et al., 1978, 90f.). Bei der Wahl der maximalen Werte für die statistischen Fehlerwahrscheinlichkeiten sollte demnach stets berücksichtigt werden, welche Konsequenzen sich aus der irrtümlichen Annahme oder Zurückweisung einer Hypothese ergeben können. Diese Überlegungen können u.U. auch zu „ungewöhnlichen“ Festlegungen wie  $\alpha = 0,20$  und  $\beta = 0,05$  führen (vgl. Abschn. 8.3).

### 11.1.3 Zur Festlegung des experimentellen Mindesteffektes EEM

Wie oben begründet, sollte der E neben  $\alpha$  und  $\beta$  auch angeben, wie groß der Experimentelle Effekt als Maß für die statistische Assoziation mindestens sein muß, um für ihn nach bestimmten inhaltlichen Kriterien „praktisch bedeutsam“ zu sein. Für dieses Ausmaß des statistischen Zusammenhangs zwischen zwei (oder mehr) Variablen sind ebenfalls Konventionen vorgeschlagen worden (Cohen, 1969, 1977): So wird beispielsweise ein  $r^2 = 0,01$  als „kleiner“, ein  $r^2 = 0,09$  als „mittlerer“ und ein  $r^2 = 0,25$  als „großer Effekt“ bezeichnet (vgl. auch Abschn. 10.2.3.2).

Da aus der Mehrzahl der derzeit vorliegenden psychologischen Hypothesen keine Aussagen über die Größe der bei ihrer Gültigkeit zu erwartenden statistischen Assoziationen abgeleitet werden können, muß gerade zu Beginn eines Forschungsprogramms zwangsläufig auf solche Konventionen zurückgegriffen werden. Später können dann die in vergleichbaren Experimenten zur Prüfung der gleichen Theorie oder Hypothese aufgedeckten Effekte zur Grundlage der Planung weiterer Untersuchungen gemacht werden (s. Abschn. 11.2).

### 11.1.4 Zur Frage der Willkür bei der Planung von Experimenten

Die eben skizzierte Planungsstrategie kann insofern kritisiert werden, als die Werte für  $\alpha$ ,  $\beta$  und EEM, die notwendigerweise vor der Datenerhebung festzulegen sind, in aller Regel stark von der Willkür des einzelnen Experimentators abhängen. Gegen diesen Einwand führt Bredenkamp (1969 b, 278f.) aus: „Die willkürlich aussehende Fixierung einer praktisch bedeutsamen Differenz und der maximalen Wahrscheinlichkeit eines Typ-II-Fehlers führt dazu, daß durch die (nicht zu begründende) Wahl des Stichprobenumfangs für jedes  $\beta$  eine Differenz implizite als bedeutsam festgelegt worden ist. . . . doch hat die explizite Festlegung einer solchen Differenz den Vorteil, daß anderen Forschern mitgeteilt werden muß, wie groß ein Mittelwertsunterschied zu sein hat, wenn die Alternativhypothese mit der Wahrscheinlichkeit  $1 - \beta$  akzeptiert werden soll.“

## 11.2 Stadium der Entscheidung über die Kausalhypothese

Wir werden nun darstellen, wie man aufgrund der Ergebnisse von Signifikanztests zu Entscheidungen über Beibehaltung oder Falsifikation wissenschaftlicher Hypothesen gelangen kann. Wir halten es für unabdingbar, diese Entscheidung nicht nur auf dem Wert der Teststatistik beruhen zu lassen, sondern dabei auch den Wert EES für die Größe des aufgedeckten experimentellen Effektes zu beachten, der sich aus den jeweils vorliegenden Daten berechnen läßt.

Wir erläutern unsere *Entscheidungsstrategie* zunächst für die Prüfung einer gerichteten statistischen Hypothese durch einen einseitigen Test (z.B. t-Test), wobei aus der wissenschaftlichen Hypothese weiterhin  $H_1$  folgen soll.<sup>33)</sup> In diesem Fall sind 4 Situationen zu unterscheiden:

(1) Die Teststatistik fällt in den von  $H_0$  abgedeckten Bereich. Bei einer  $H_1 : \mu_1 - \mu_2 > 0$  heißt das beispielsweise, daß die Teststatistik  $t$  einen Wert kleiner oder gleich Null annimmt. Wir nehmen dann die  $H_0$  an. Sollten derartige Ergebnisse wiederholt auftreten, kann man sich zur Falsifikation der wissenschaftlichen Hypothese entschließen.

(2) Die Teststatistik fällt in den Rejektionsbereich, und der EES ist größer als der festgelegte Wert EEM. In diesem Fall wird die  $H_1$  akzeptiert, und die WH hat sich (vorläufig) bewährt.

(3) Der Wert der Teststatistik fällt in den Rejektionsbereich, EES ist jedoch geringer als EEM. Wie im Fall (2) wird die  $H_1$  angenommen, denn die Wahrscheinlichkeit, daß dies irrtümlich geschieht, ist höchstens gleich  $\alpha$ , also einem (sehr) geringen Wert. Der Ausgang des Experiments spricht grundsätzlich für die wissenschaftliche Hypothese; allerdings sollte in einem Folgeexperiment untersucht werden, ob der kleinere EE replizierbar ist. Ist dies nicht oder nur sporadisch der Fall, sollte man die WH als schlecht bewährt ansehen. Auf die andere Möglichkeit, daß der geringere Effekt wiederholbar ist, gehen wir im Anschluß an Situation (4) ein.

(4) Der Wert der Teststatistik weist auf eine Abweichung von der  $H_0$  in der von  $H_1$  spezifizierten Richtung hin, ist aber zu klein, um zu einem signifikanten Ergebnis zu führen. Nach unserer Entscheidungsstrategie im Signifikanztest nehmen wir dann die  $H_0$  an; das bedeutet, daß der experimentelle Effekt als zufällige Abweichung interpretiert wird. Dies spricht zunächst einmal gegen die Kausalhypothese.

Weitere Prüfungen sollten mit einem kleineren EEM, einer größeren Zahl von Versuchspersonen und dadurch mit einer höheren Präzision durchgeführt werden. Wird dabei insgesamt das zuerst erhaltene Ergebnis repliziert, muß das aber nicht unbedingt zur Falsifikation der wissenschaftlichen Hypothese führen. Ergeben sich nämlich mit einer gewissen Regelmäßigkeit nur sehr kleine Abweichungen von der Nullhypothese in Richtung auf die Alternativ-

---

<sup>33)</sup> Wir erweitern und modifizieren damit den Gedanken eines dreigeteilten Stichprobenraums, der erstmals von Neyman & Pearson (1933 b, 493) entwickelt wurde und der dann mit unterschiedlichen Akzentuierungen nacheinander von Lehmann (1950, 23f.), Kaiser (1960), Meehl (1967), Kleiter (1969), Bredenkamp (1972, 80-84) und Bortz, Österreich & Vogelbusch (1979) wieder aufgegriffen und weiterentwickelt worden ist (zur ausführlichen Darstellung siehe Westermann & Hager, 1982).

hypothese, kann der E zwischen zwei Möglichkeiten entscheiden, und zwar unabhängig davon, ob die Abweichungen überwiegend „signifikant“ sind (obiger Fall 3) oder nicht (Fall 4): Er kann die wissenschaftliche Hypothese falsifizieren, weil sie zu wenig an Datenvarianz aufklärt (Bredenkamp, 1972, 83-84), oder er behält sie bei, allerdings mit der Einschränkung, daß ihr nur geringe statistische Assoziationen entsprechen.

Bei dieser Entscheidung ist auch zu berücksichtigen, inwieweit für die durchgeführten Untersuchungen die besprochenen Aspekte der experimentellen Validität gegeben sind. So kann es z.B. von besonderem Interesse sein, zu prüfen, inwieweit sich die Effekte auch unter „natürlichen“ Bedingungen (Feldexperimente, nicht-reaktive Operationalisierungen usw.) zeigen.

Ferner ist bei Falsifikationsentscheidungen stets auch zu berücksichtigen, ob eine alternative Hypothese zur Erklärung der empirischen Ergebnisse gefunden werden kann. Damit derartige Entscheidungen nachvollziehbar und auch kritisierbar werden, ist es unbedingt erforderlich, für jedes Experiment als spezielles und wichtiges Ergebnis auch die Größe des aufgedeckten experimentellen Effektes mitzuteilen.

Muß die wissenschaftliche Hypothese über einen *zweiseitigen Test* geprüft werden, tritt zusätzlich folgendes Problem auf: Die Nullhypothese (z. B.  $\mu = 0$ ) definiert keine Parametermenge, die eindeutig für  $H_0$  spricht, sondern nur genau *einen* Wert. Da die Signifikanztests in der Regel von kontinuierlichen Variablen ausgehen, ist dann die Wahrscheinlichkeit, einen Wert der Teststatistik zu erhalten, der genau dem in  $H_0$  spezifizierten Wert entspricht, gleich Null. Eine Annahme von  $H_0$  (und damit eine Ablehnung von  $H_1$ ) ist deshalb bei zweiseitigen Tests nur möglich, wenn man um den in  $H_0$  spezifizierten Parameterwert eine Menge von Stichprobenergebnissen festlegt, die - falls sie auftreten - auch noch für  $H_0$  sprechen sollen. Beim hier vorgeschlagenen Signifikanztest in der Neyman-Pearson-Tradition erfolgt dies durch eine vorherige Festlegung von EEM,  $\alpha$  und  $\beta$ . Dadurch wird ein Bereich von möglichen Werten der Stichprobenstatistik (oder äquivalent der Effektgröße) festgelegt, der zur Annahme von  $H_0$  führt. Dieser Bereich liegt zwischen dem unteren und dem oberen kritischen Wert der Teststatistik und entspricht Effektgrößen zwischen Null und einem Wert EEK, der bei adäquat fixierter Teststärke notwendigerweise kleiner ist als der gewählte EEM.

Wir unterscheiden beim zweiseitigen Signifikanztest drei Entscheidungssituationen:<sup>34)</sup>

<sup>34)</sup> Der Einfachheit halber gelte weiterhin, daß aus der wissenschaftlichen Hypothese die statistische Alternativhypothese folgt, obwohl dieser Fall weniger häufig auftritt (vgl. Abschn. 8.1.1, Punkt 2). Die hier dargestellten Entscheidungsstrategien sind aber problemlos auf den Fall der Implikationsbeziehung  $WH \rightarrow H_0$  übertragbar; vgl. dazu Westermann & Hager (1982).

(1) Der Test führt nicht zu einem signifikanten Ergebnis, EES ist also kleiner als EEK.  $H_0$  ist in diesem Falle anzunehmen, und bei zuverlässigen Resultaten dieser Art kann die WH falsifiziert werden.

(2) Der empirische Wert der Teststatistik fällt in den Rejektionsbereich, und EES ist nicht kleiner als EEM. Es wird die  $H_1$  angenommen, und die WH hat sich (vorläufig) bewährt.

(3) Der empirische Wert der Teststatistik fällt zwar in den Rejektionsbereich, aber EES ist kleiner als EEM. Nach den Entscheidungsregeln des Signifikanztests muß auch in diesem Fall die  $H_1$  akzeptiert werden. Allerdings kann die bei der Prüfung gerichteter Hypothesen ausgesprochene Empfehlung, weitere Tests mit einem entsprechend verringerten Wert für EEM durchzuführen, in dieser Situation nicht wiederholt werden, weil dadurch der für  $H_0$  sprechende Parameterbereich zwangsläufig ebenfalls verringert würde. Eine derartige Entscheidung kann vom Experimentator nur fallspezifisch und gut begründet getroffen werden. Treten nun in weiteren Hypothesenprüfungen überwiegend experimentelle Effekte auf, die zu signifikanten Abweichungen von der  $H_0$  führen, aber kleiner als EEM sind, stellt sich die Frage nach der Interpretation dieser Befunde. Wie im Falle der einseitigen Tests kann diese Frage auch hier nicht nach allgemeinen Kriterien beantwortet werden: Man kann sich entscheiden, die wissenschaftliche Hypothese mit der Einschränkung als bewährt anzusehen, daß sie nur (relativ) geringen statistischen Assoziationen entspricht, oder aber diese (zu) geringen Effekte als Basis für eine Falsifikation akzeptieren.

Prinzipiell die gleichen Überlegungen wie zu zweiseitigen Tests gelten auch bei Prüfungen statistischer Hypothesen über Verfahren, bei denen die Nullhypothese zwar einseitig getestet wird, mit ihrer Gültigkeit aber auch nur ein einziger Wert vereinbar ist. Beim (einseitigen) Test der (ungerichteten) Nullhypothese der Varianzanalyse über die F-Verteilung beispielsweise ist dieser Wert  $F = df_N/(df_N - 2)$  (Hays, 1977, 445). Zu Einzelheiten siehe Westermann & Hager (1982).

Insgesamt soll eine Entscheidung über die Bewährung oder Falsifikation einer wissenschaftlichen Hypothese aufgrund der Prüfung einer statistischen Hypothese also nie allein auf der Signifikanz oder Nicht-Signifikanz des Stichprobenergebnisses beruhen, sondern stets auch auf dem erhaltenen experimentellen Effekt, der ja ein Maß für die Stärke der statistischen Assoziation zwischen UV(n) und AV(n) darstellt. Eine Angabe über die Größe des EES sollte daher in Forschungsberichten *stets* enthalten sein, und zwar aus den im Abschnitt 9.5 genannten Gründen sowohl in absoluten wie in relativen oder normierten Einheiten. Diese Angabe ist besonders wichtig, wenn der Experimentator seine wissenschaftliche Hypothese beibehalten hat, obwohl die empirisch festgestellten statistischen Assoziationen nur gering waren.

Die routinemäßige Angabe der aufgedeckten experimentellen Effekte würde es auch erleichtern, die Ergebnisse mehrerer Untersuchungen zur Prüfung der gleichen Hypothese zusammenzufassen: Man könnte einen mittleren Wert für die statistische Assoziation berechnen, die empirisch zwischen den durch die Hypothese in Zusammenhang gebrachten Variablen bestehen (Bredenkamp, 1972, 83; Glass, 1978; Pillemer & Light, 1980; s.a. Abschn. 1.3). Des weiteren könnten nach unterschiedlichen Kriterien (Art der Operationalisierung, Population, Situation usw.) gebildete Teilmengen der vorliegenden Untersuchungen auf diese Weise zusammengefaßt werden. Auf dieser breiten Basis wäre dann zu entscheiden, ob die wissenschaftliche Kausalhypothese sich bewährt hat, ob sie falsifiziert werden sollte oder ob ihr Geltungsbereich einzuschränken ist.

## *Literatur*

- Abrahams, N. M. & Alf, E. F. Jr. 1978. Relative costs and statistical power in the extreme groups approach. *psychometrika*, 43, 11-17.
- Acock, A. C. & Stavig, G. R. 1979. A measure of association for nonparametric statistics. *Social Forces*, 57, 1381-1386.
- Adam, J. & Enke, H. 1972. Analyse mehrdimensionaler Kontingenztafeln mit Hilfe des Informationsmaßes von Kullback. *Biometrische Zeitschrift*, 14, 305-323.
- Adams, E. R., Fagot, R. F. & Robinson, R. E. 1965. A theory of appropriate statistics. *Psychometrika*, 30, 99-127.
- Alf, E. F. Jr. & Abrahams, N. M. 1972. Comment on component randomization tests. *Psychological Bulletin*, 77, 223-224.
- Alf, E. F. Jr. & Abrahams, N. M. 1973. Reply to Edgington. *Psychological Bulletin*, 80, 86-87.
- Alimena, B. S. 1962. A method of determining unbiased distribution in the latin square. *Psychometrika*, 27, 315-317.
- Allerbeck, K. R. 1978. Meßniveau und Analyseverfahren: Das Problem „strittiger“ Intervallskalen. *Zeitschrift für Soziologie*, 7, 199-214.
- Amir, Y. 1969. Contact hypothesis in ethnic relations. *Psychological Bulletin*, 71, 319-342.
- Anderson, N. H. 1961. Scales and statistics: Parametric and non-parametric. *Psychological Bulletin*, 58, 305-316.
- Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. & Tukey, J. W.: 1972. Robust estimates of location: Survey and advances. Princeton: Princeton University Press.
- Ancombe, F. J. & Tukey, J. W. 1963. The examination and analysis of residuals. *Technometrics*, 5, 141-160.

- Armitage, P. & Remington, R. D. 1970. Experimental design. In: McArthur, J. W. & Colton, Th. (Eds): Statistics in endocrinology. Cambridge, Mass.: The MIT Press, 3-31.
- Arnold, W. (Ed.) 1972'. Psychologisches Praktikum. Band 1. Stuttgart: Fischer.
- Aronson, E. & Carlsmith, J. M. 1968<sup>2</sup>. Experimentation in socialpsychology. In: Lindzey, G. & Aronson, E. (Eds): Handbook of social psychology. Vol. II. Reading, Mass.: Addison-Wesley, 1-79.
- Aspin, A. A. 1948. An examination and further development of a formula occuring in the problem of comparing two mean values. Biometrika, 35, 88-96.
- Aspin, A. A. 1949. Tables for use in comparisons whose accuracy involves two variances separately estimated. Biometrika, 36, 290-296.
- Auslitz, K., Hesse, H. G. & Rieder, A. 1975. Möglichkeiten und Probleme der Anwendung des Allgemeinen Linearen Modells auf Fragestellungen der Varianz-, Regressions- und Kovarianzanalyse. Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung und Gesellschaft zur Förderung Pädagogischer Forschung, Mitteilungen und Nachrichten, 79/80, 1-78.
- Bailey, D. E. 1971. Probability and Statistics. Modells for Research. New York: Wiley.
- Bakan, D. 1966. The test of significance in psychological research. Psychological Bulletin, 66, 423-437.
- Baker, B. O., Hardych, C. D. & Petrionovich, L. F. 1966. Weak measurements vs. strong statistics: An empirical critique to S. S. Stevens proscription on statistics. Educational and Psychological Measurement, 26, 291-309.
- Baker, F. B. & Collier, R. O. Jr. 1966. Some empirical results on variance ratios under permutation in the completely randomized design. Journal of the American Statistical Association, 63, 813-820.
- Baker, F. B. & Collier, R. O. Jr. 1968. An empirical study into factors affecting the F-test under Permutation for the randomized block design. Journal of the American Statistical Association, 63, 902-911.
- Barber, T. X. 1976. Pitfalls in human research. New York: Pergamon.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D. 1972. Statistical inference under order restrictions. New York: Wiley.
- Barnett, V. 1973. Comparative statistical inference. London: Wiley.
- Barnett, V. & Lewis, T. 1978. Outliers in statistical data. Chichester: Wiley.
- Bartenwerfer, H. & Raatz, U. 1979. Einführung in die Psychologie, Bd. 6: Methoden der Psychologie. Wiesbaden: Akademische Verlagsgesellschaft.
- Bartholomew, D. J. 1959 (a). A test of homogeneity for ordered alternatives. Biometrika, 46, 36-48.
- Bartholomew, D. J. 1959 (b). A test of homogeneity for ordered alternatives II. Biometrika, 46, 328-335.
- Bartholomew, D. J. 1961 (a). Ordered tests in the analysis of variance. Biometrika, 48, 325-332.



- Bartholomew, D. J. 1961 (b). A test of homogeneity of means under restricted alternatives. *Journal of the Royal Statistical Society, Series B*, 23, 239-281.
- Bartlett, M. S. 1937. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society, Series A*, 160, 268-282.
- Bartlett, M. S. 1939. A note on tests of significance in multivariate analysis. *Proceedings of the Cambridge Philosophical Society*, 35, 180-185.
- Bartlett, M. S. 1947. The use of transformations. *Biometrics*, 3, 39-52.
- Beauchamp, K. B. & May, R. B. 1964. Replication report: Interpretation of levels of significance by psychological researchers. *Psychological Reports*, 14, 272.
- Bauknecht, K. Kohlas, J. & Zehnder, C. A. 1976. *Simulationstechnik*. Berlin: Springer.
- Behnken, D. W. & Draper, N. R. 1972. Residuals and their variance Patterns. *Technometrics*, 14, 101-111.
- Behrens, W. V. 1929. Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher*, 68, 807-837.
- Bern, D.J. 1972. Self-perception theory. In: Berkowitz, L. (Ed.): *Advances in experimental social psychology*. Vol. 6. New York: Academic Press, 1-66.
- Bern, D. J. 1975. *Meinungen, Einstellungen, Vorurteile*. Köln: Bensinger & Sauerländer.
- Benninghaus, H. 1974, 1979<sup>3</sup>. *Deskriptive Statistik*. Stuttgart: Teubner.
- Berchtold, H. 1979. A modified Mann-Whitney test with improved asymptotic relative efficiency. *Biometrical Journal*, 21, 649-655.
- Berenson, M. L. 1978. A simple distribution-free test for trend in one-way layouts. *Educational and Psychological Measurement*, 38, 905-912.
- Berg, J. A. (Ed.) 1967. *Response set in personality measurement*. Chicago: Aldine.
- Bernhardson, C. S. 1975. Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics*, 31, 229-232.
- Betz, M. A. & Gabriel, K. R. 1978. Type IV errors and analysis of simple effects. *Journal of Educational Statistics*, 3, 121-143.
- Bevan, M. F., Benton, J. Q. & Myers, J. L. 1974. The robustness of the F-test to violations of continuity and form of treatment population. *British Journal of Mathematical and Statistical Psychology*, 27, 199-204.
- Bickel, P. J. 1976. Another look at robustness: A review of reviews and some developments. *Scandinavian Journal of Statistics*, 3, 145-168.
- Bishop, T. A. 1978. A Stein two-sample procedure for the general model with unequal variances. *Communications in Statistics, A* 7(5), 495-507.
- Blair, R. C. & Higgins, J. J. 1978. Tests of hypotheses for unbalanced factorial designs under various regression/coding method combinations. *Educational and Psychological Measurement*, 38, 621-631.

- Blair, R. C., Higgins, J. J. & Smitley, W. D. S. 1980. On the relative power of the U and t test. *British Journal of Mathematical and Statistical Psychology*, 33, 114-120.
- Bliss, C. J. 1967, 1970. *Statistics in biology*. Vol. 1: 1967. Vol. 2: 1970. New York: McGraw-Hill.
- Bock, R. D. 1963, 1967. Multivariate analysis of variance of repeated measurements. In: Harris, C. W. (Ed.): *Problems in measuring change*. Madison: The University of Wisconsin Press, 85-103.
- Bock, R. D. 1975. *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- Bock, R. D. & Haggard, E. A. 1968. The use of multivariate analysis of variance in behavioral research. In: Whitla, D. K. (Ed.): *Handbook of measurement and assessment in behavioral sciences*. Reading, Mass.: Addison-Wesley, 100-142.
- Boersma, F. D., de Jonge, J. J. & Stellwagen, W. R. 1964. A power comparison of the F and L tests - I. *Psychological Review*, 71, 505-513.
- Boik, R. J. 1979. The rationale of Scheffé's method and the simultaneous test procedure. *Educational and Psychological Measurement*, 39, 49-56.
- Bolles, R. C. & Messick, S. 1958. Statistical Utility in experimental inference. *Psychological Reports*, 4, 223-227.
- Boneau, C. A. 1962. A comparison of the power of the U and t tests. *Psychological Review*, 69, 246-256.
- Borich, G. O. & Godbout, R. C. 1974. Extreme groups design and calculation of statistical power. *Educational and Psychological Measurement*, 34, 663-675.
- Bortz, J. 1977, 1979<sup>2</sup>. *Lehrbuch der Statistik für Sozialwissenschaftler*. Berlin: Springer.
- Bortz, J., Österreich, R. & Vogelbusch, W. 1979. Die Ermittlung optimaler Stichprobenumfänge für die Durchführung von Binomial-Tests. *Archiv für Psychologie*, 131, 267-292.
- Bowman, K. O., Beauchamp, J. J. & Shenton, L. R. 1977. The distribution of the t-statistic under non-normality. *International Statistical Review*, 45, 233-242, 256.
- Box, G. E. P. 1950. Problems in the analysis of growth and wear curves. *Biometrics*, 6, 362-389.
- Box, G. E. P. 1953. Non-normality and tests on variances. *Biometrika*, 40, 318-335.
- Box, G. E. P. 1954 (a). Some theorems on quadratic forms applied in the study of analysis of variance problems. I: Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 25, 290-302.
- Box, G. E. P. 1954 (b). Some theorems on quadratic forms applied in the study of analysis of variance problems. II: Effect of inequality of variance and correlation of errors in the two-way-classification. *Annals of Mathematical Statistics*, 25, 484-498.

- Box, G. E. P. 1968. Experimental design: Response surfaces. In: Sills, D. L. (Ed.): International Encyclopedia of the Social Sciences. Vol. 5. Boston: Crowell Collier and Macmillan, 259-259.
- Box, G. E. P. & Cox, D. R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211-252.
- Box, G. E. P., Hunter, W. G. Jr. & Hunter, J. S. 1978. *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York: Wiley-Interscience.
- Box, G. E. P. & Jenkins, G. M. 1970, 1976<sup>2</sup>. *Time series analysis, forecasting and control*. San Francisco: Holden-Day.
- Box, G. E. P. & Tiao, G. C. 1973. *Bayesian inference in statistical analysis*. Reading, Mass.: Addison-Wesley.
- Bozarth, J. D. & Roberts, R. R. Jr. 1972. Signifying significant significance. *American Psychologist*, 27, 774-775.
- Bracht, G. H. 1970. Experimental factors related to aptitude-treatment interactions. *Review of Educational Research*, 40, 627-645.
- Bracht, G. H. & Glass, G. V. 1968. The external validity of experiments. *American Educational Research Journal*, 5, 437-474.
- Bradley, J. V. 1968. *Distribution-free statistical tests*. Englewood Cliffs, N. J.: Prentice-Hall.
- Bradley, J. V. 1972. Nonparametric statistics. In: Kirk, R. E. (Ed.): *Statistical issues*. Belmont, Cal.: Wadsworth, 329-338.
- Bradley, J. V. 1977. A common Situation conducive to bizarre distribution shapes. *The American Statistician*, 31, 147-150.
- Bradley, J. V. 1978. Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Bradley, J. V. 1980 (a). Nonrobustness in one-sample Z and t tests: a large scale sampling study. *Bulletin of the Psychonomic Society*, 15, 29-32.
- Bradley, J. V. 1980 (b). Nonrobustness in classical tests on means and variances: A large-scale sampling study. *Bulletin of the Psychonomic Society*, 15, 275-278.
- Bradley, J. V. 1980 (c). Nonrobustness in Z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society*, 16, 333-336.
- Brand, M. 1976. *The nature of causation*. Urbana: University of Illinois Press.
- Bredenkamp, J. 1968. F-Tests zur Prüfung von Trends und Trendunterschieden. *Zeitschrift für experimentelle und angewandte Psychologie*, 15, 239-272.
- Bredenkamp, J. 1969 (a). Experiment und Feldexperiment. In: Graumann, C. F. (Hrsg.): *Handbuch der Psychologie*. Bd. 7: Sozialpsychologie. I. Halbband: Theorien und Methoden. Göttingen: Hogrefe, 332-374.
- Bredenkamp, J. 1969 (b). über die Anwendung von Signifikanztests bei Theorie-testenden Experimenten. *Psychologische Beiträge*, 11, 275-285.

- Bredenkamp, J. 1970. Über Maße der praktischen Signifikanz. Zeitschrift für Psychologie, 177, 310-318.
- Bredenkamp, J. 1971. Bemerkungen zu den Trendanalysen nach Ferguson. Zeitschrift für experimentelle und angewandte Psychologie, 18, 386-391.
- Bredenkamp, J. 1972. Der Signifikanztest in der psychologischen Forschung. Frankfurt am Main: Akademische Verlagsgesellschaft.
- Bredenkamp, J. 1974. Nonparametrische Prüfung von Wechselwirkungen. Psychologische Beiträge, 16, 398-416.
- Bredenkamp, J. 1975. Varianzanalytische und regressionsanalytische Verfahren in der Curriculumevaluation. In: Frey, K. (Hrsg.): Curriculum-Handbuch. München: Piper, 786-822.
- Bredenkamp, J. 1979. Das Problem der externen Validität pädagogisch-psychologischer Untersuchungen. In: Brandstätter, J., Reinert, G. & Schneewind, K. A. (Hrsg.): Pädagogische Psychologie. Probleme und Perspektiven. Stuttgart: Klett-Cotta, 267-289.
- Bredenkamp, J. 1980. Theorie und Planung psychologischer Experimente. Darmstadt: Steinkopff.
- Bredenkampf, J. & Feger, H. 1970. Kriterien für die Entscheidung über die Aufnahme empirischer Arbeiten in die Zeitschrift für Sozialpsychologie. Zeitschrift für Sozialpsychologie, 1, 43-47.
- Bredenkamp, J. & Hager, W. 1979. Experimentelle Befunde zur modifizierten Invarianzhypothese und zur Hypothese einer konstanten Langzeitgedächtnisspanne. Psychologische Beiträge, 21, 382-400.
- Breen, L. & Gaito, J. 1970. Comments on Friedman's  $r_m$  procedure. Psychological Bulletin, 73, 309-310.
- Brewer, J. K. 1972. On the power of statistical tests in the American Educational Research Journal. American Educational Research Journal, 9, 391-401.
- Bridgman, P. W. 1927. The logic of modern physics. New York: Macmillan.
- Brillinger, D. R. 1975. Time series, data analysis and theory. New York: Holt, Rinehart & Winston.
- Brocke, B. 1978. Technologische Prognosen. Freiburg: Alber.
- Brocke, B. 1979. Aspekte einer Methodologie der angewandten Sozial- und Verhaltenswissenschaften. Zeitschrift für Sozialpsychologie, 10, 2-29.
- Broekmann, N. C. 1973. Note on the influence of an exceptional subject. Psychometrika, 38, 611-613.
- Brown, D. J. 1975. Down with the linear model. American Educational Research Journal, 491-505.
- Brown, M. B. & Forsythe, A. B. 1974 (a). The small sample behavior of some statistics which test the equality of several means. Technometrics, 16, 129-132.
- Brown, M. B. & Forsythe, A. B. 1974 (b). The ANOVA and multiple comparisons for the data with heterogeneous variances. Biometrics, 30, 719-724.

- Brown, M. B. & Forsythe, A. B. 1974 (c). Robust tests for equality of variances. *Journal of the American Statistical Association*, 69, 364-367.
- Büning, H. & Trenkler, G. 1978. *Nichtparametrische statistische Methoden*. Berlin: de Gruyter.
- Bugelski, B. R. 1960. *A first course in experimental psychology*. New York: Holt.
- Bungard, W. (Hrsg.) 1980. *Die „gute“ Versuchsperson denkt nicht*. München: Urban und Schwarzenberg.
- Bungard, W. & Lück, H. E. 1974. *Forschungsartefakte und nicht-reaktive Meßverfahren*. Stuttgart: Teubner.
- Bunge, M. 1967. (a, b). *Scientific Research*. I. The search for system. 1967 (a). II. The search for truth. 1967 (b). New York: Springer.
- Burnett, T. D. & Barr, D. R. 1977. A non-metric analogy of analysis of covariance. *Educational and Psychological Measurement*, 37, 341-348.
- Busemeyer, J. R. 1980. Importance of measurement theory, error theory, and experimental design for testing the significance of interactions. *Psychological Bulletin*, 88, 237-244.
- Callaway, J. N., Nowicki, S. & Duke, M. P. 1980. Overt expression of experimental expectancies, interaction with subject expectancies and Performance in a psychomotor task. *Journal of Research in Personality*, 14, 27-39.
- Campbell, D. T. 1957. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297-312.
- Campbell, D. T. 1969. Prospective: Artifact and control. In: Rosenthal, R. L. & Rosnow, R. C. (Eds): *Artifact in behavioral research*. New York: Academic Press, 351-382.
- Campbell, D. T. & Fiske, D. W. 1959. Convergent and discriminant Validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Campbell, D. T. & Stanley, J. C. 1963, 1973. Experimental and quasi-experimental designs for research in teaching. In: Gage, N. L. (Ed.): *Handbook of research in teaching*. Chicago: Rand McNally, Kap. 5. Deutsche Bearbeitung in: Ingenkamp, K. (Hrsg.): *Strategien der Unterrichtsforschung*. Weinheim: Beltz, 101-193.
- Carlsmith, J. M., Ellsworth, P. C. & Aronson, E. 1976. *Methods of research in social psychology*. Reading, Mass. : Addison-Wesley.
- Carlson, R. 1976. Discussion: The logic of tests of significance. *Philosophy of Science*, 43, 116-128.
- Carmer, S. G. & Swanson, M. R. 1971. Detection of differences between means: A Monte Carlo study of five pairwise multiple comparison procedures. *Agronomy Journal*, 63, 940-945.
- Carmer, S. G. & Swanson, M. R. 1973. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association*, 68, 66-74.

- Carnap, R. 1956, 1960. The methodological Character of theoretical concepts. In: Feigl, H. & Scriven, M. (Eds): Minnesota studies in the philosophy of science. Vol. I. Minneapolis: University of Minnesota Press, 33-76. Deutsche Übersetzung: Theoretische Begriffe der Wissenschaft: Eine logische und methodologische Untersuchung. Zeitschrift für philosophische Forschung, 14, 209-233 und 571-598.
- Carroll, R. M. & Nordholm, L. A. 1975. Sampling characteristics of Kelley's  $\epsilon$  and Hays'  $\omega^2$ . Educational and Psychological Measurement, 35, 541-554.
- Carter, D. S. 1979. Comparison of different shrinkage formulars in estimating population multiple correlation coefficients. Educational and Psychological Measurement, 39, 261-266.
- Carver, R. P. 1968. Note on a schema for proper utilization of multiple comparisons in research. American Educational Research Journal, 5, 730-732.
- Carver, R. P. 1978. The case against statistical significance testing. Harvard Educational Review, 48, 378-399.
- Cascio, W. F., Valenzi, E. R., Silbey, V. 1978. Validation and statistical power: Implications for applied research. Journal of Applied Psychology, 63, 589-595.
- Cascio, W. F., Valenzi, E. R. & Silbey, V. 1980. More on Validation and statistical power. Journal of Applied Psychology, 65, 135-138.
- Chase, L. J. & Baran, S. J. 1976. An assessment of quantitative research in mass communication. Journalism Quarterly, 53, 308-311.
- Chase, L.J. & Chase, R. B. 1976. A statistical power analysis of applied psychological research. Journal of Applied Psychology, 61, 234-237.
- Chase, L. J. & Tucker, R. K. 1975. A power-analytic examination of contemporary communication research. Speech Monographs, 42, 29-41.
- Chase, L. J. & Tucker, R. K. 1976. Statistical power: Derivation, development and data-analytic implications. Psychological Record, 26, 473-486.
- Chatfield, C. 1975. The analysis of time series: Theory and practice. London: Chapman & Hall.
- Church, J. D. & Wike, E. L. 1976. The robustness of homogeneity of variance tests for asymmetric distributions: A Monte Carlo study. Bulletin of the Psychonomic Society, 7, 416-420.
- Church, J. D. & Wike, E. L. 1979. A Monte Carlo study of nonparametric multiple-comparison tests for a two-way layout. Bulletin of the Psychonomic Society, 14, 95-98.
- Church, J. D. & Wike, E. L. 1980. Two Monte Carlo studies of Silverstein's nonparametric multiple comparison tests. Psychological Reports, 46, 403-407.
- Clauß, G. & Ebner, H. 1978<sup>6</sup>. Grundlagen der Statistik. Berlin: Volk und Wissen.
- Cleary, T. A. & Linn, R. L. 1969. Error of measurement and the power of a statistical test. British Journal of Mathematical and Statistical Psychology, 22, 49-55.

- Cleary, T. A., Linn, R. L. & Walster, G. W. 1970. Effect of reliability and validity on power of statistical tests. In: Borgatta, E. F. & Bohrnstedt, G. W. (Eds): *Sociological methodology*. 1970. San Francisco: Jossey-Bass, 130-138.
- Cochran, W. G. 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 3, 22-38.
- Cochran, W. G. 1951. Testing a linear relation among variances. *Biometrics*, 7, 17-32.
- Cochran, W. G. 1957. Analysis of covariance: Its nature and uses. *Biometrics*, 13, 261-281.
- Cochran, W. G. 1968 (a). The design of experiments. In: Sills, D. L. (Ed.): *International Encyclopedia of the Social Sciences*. Vol. 5. Boston: Crowell Collier and Macmillan, 245-254.
- Cochran, W. G. 1968 (b). Errors of measurement in statistics. *Technometrics*, 10, 637-666.
- Cochran, W. G. 1970. Some effects of errors of measurement on multiple correlation. *Journal of the American Statistical Association*, 65, 22-34.
- Cochran, W. G. 1977<sup>3</sup>, 1972. *Sampling Techniques*. New York: John Wiley. Deutsch: Stichprobenverfahren. Berlin: De Gruyter.
- Cochran, W. G. & Bliss, C. I. 1970. Analysis of variance. In: McArthur, J. W. & Colton, Th. (Eds): *Statistics in endocrinology*. Cambridge, Mass.: The MIT Press, 33-78.
- Cochran, W. G. & Cox, G. M. 1957<sup>2</sup>, 1968<sup>3</sup>. *Experimental designs*. New York: Wiley.
- Cohen, J. 1962. The statistical power of abnormal - social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.
- Cohen, J. 1965. Some statistical issues in psychological research. In: Wolman, B. B. (Ed.): *Handbook of clinical psychology*. New York: MacGraw-Hill, 95-121.
- Cohen, J. 1968. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70, 426-443.
- Cohen, J. 1969, 1977<sup>2</sup>. *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Cohen, J. 1970. Approximate power and sample size determination for common one-sample and two-sample hypothesis tests. *Educational and Psychological Measurement*, 30, 811-831.
- Cohen, J. 1973 (a). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, 33, 107-112.
- Cohen, J. 1973 (b). Statistical power analysis and research results. *American Educational Research Journal*, 10, 225-230.
- Cohen, J. 1980. Trend analysis the easy way. *Educational and Psychological Measurement*, 40, 565-568.
- Cohen, J. & Cohen, P. 1975. *Applied multiple regression/correlation analysis for the Behavioral Sciences*. Hillsdale, N. J.: Lawrence Erlbaum.

- Collier, R. O. Jr. & Baker, F. B. 1966. Some Monte Carlo results on the power of the F-test under permutation in the simple randomized block design. *Biometrika*, 53, 199-203.
- Collier, R. O. Jr., Baker, F. B., Mandeville, G. K. & Hayes, T. F. 1967. Estimates of test size for several test procedures based on conventional variance ratio in the repeated measures design. *Psychometrika*, 32, 339-353.
- Collier, R. O. Jr. & Larson, R. C. 1969. Monte Carlo methods. *Review of Educational Research*, 39, 734-739.
- Cook, S. W. & Selltitz, C. 1964. A multiple-indicator approach to attitude measurement. *Psychological Bulletin*, 62, 36-55.
- Cook, T. D. & Campbell, D. T. 1976. The design and conduct of quasi-experiments and true experiments in field settings. In: Dunnette, M. D. (Ed.): *Handbook of industrial and organizational psychology*. Chicago: Rand McNally, 223-326.
- Cook, T. D. & Campbell, D. T. 1979. *Quasi-experimentation. Design and analysis issues for field settings*. Chicago: Rand McNally.
- Cooley, W. W. & Lohnes, P. R. 1971. *Multivariate data analysis*. New York: Wiley.
- Coombs, C. H., Dawes, R. M. & Tversky, A. 1975. *Mathematische Psychologie*. Weinheim: Beltz.
- Cooper, H. M. & Rosenthal, R. 1980. Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 87, 442-449.
- Cowles, M. P. 1974. N = 35: A rule of thumb for psychological researchers. *Perceptual and Motor Skills*, 38, 1135-1138.
- Cox, D. R. 1957. The use of a concomitant variable in selecting an experimental design. *Biometrika*, 44, 150-158.
- Cox, D. R. 1958. *Planning of experiments*. New York: Wiley.
- Cox, D. R. 1961. The design of experiments: The control of error. *Journal of the Royal Statistical Society, Series A*, 124, 44-48.
- Craig, J. R., Eison, C. L., Metzger, L. P. 1976. Significance tests and their Interpretation: An example utilizing published research and  $\omega^2$ . *Bulletin of the Psychonomic Society*, 7, 280-282.
- Cramer, E. M. & Appelbaum, M. I. 1980. Nonorthogonal Analysis of Variance - once again. *Psychological Bulletin*, 87(1), 51-57.
- Cramer, E. M. & Bock, R. R. 1966. Multivariate analysis. *Review of Educational Research*, 36, 604-617.
- Cramer, E. M. & Nicewander, W. A. 1979. Some symmetric, invariant measures of multivariate association. *Psychometrika*, 44, 43-54.
- Cronbach, L. J. 1957. The two disciplines of scientific psychology. *American Psychologist*, 12, 671-689.
- Cronbach, L. J. 1975. Beyond the two disciplines of scientific psychology. *American Psychologist*, 30, 116-127.



- Cronbach, L. J. & Furby, L. 1970. How should we measure „change“ - or should we? *Psychological Bulletin*, 74, 68-80.
- Cronbach, L. J. & Meehl, P. E. 1955. Construct Validation in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Cureton, E. E. 1966. On correlation coefficients. *Psychometrika*, 31, 605-607.
- D'Agostino, R. B. 1972. Relation between the chi-squared and ANOVA tests for testing the equality of k independent dichotomous populations. *American Statistician*, 26(3), 30-32.
- Darlington, R. B. 1968. Multiple regression in psychological research and practice. *Psychological Bulletin*, 69, 161-182.
- David, H. A., Lachenbruch, P. A. & Brandis, H. P. 1972. The power function of range and studentized range tests in normal samples. *Biometrika*, 59, 161-168.
- Davidson, M. L. 1972. Univariate versus multivariate tests in repeated measures experiments. *Psychological Bulletin*, 77, 446-452.
- David, D. J. 1969. Flexibility and power in comparisons among means. *Psychological Bulletin*, 71, 441-444.
- Dawes, R. M. 1977. *Grundlagen der Einstellungsmessung*. Weinheim: Beltz.
- Dayton, C. M. & Schafer, W. D. 1973. Extended tables of t and Chi Square for Bonferroni tests with unequal error allocation. *Journal of the American Statistical Association*, 68, 78-83.
- Dénes, J. & Keedwell, A. D. 1974. *Latin squares and their applications*. Budapest: Akademiai Kiado.
- Derrick, T. 1976. The criticism of inferential statistics. *Educational Research*, 19(1), 35-40.
- Diehl, J. M. 1977, 1979<sup>3</sup>. *Varianzanalyse*. Frankfurt am Main: Fachbuchhandlung für Psychologie.
- Dierkes, M. 1977. Die Analyse von Zeitreihen und Longitudinalstudien. In: Koolwijk, J. van & Wieken-Mayser, M. (Hrsg.): *Techniken der empirischen Sozialforschung*. Bd. 7. München: Oldenbourg, 111-169.
- Digman, J. M. 1966. Interaction and nonlinearity in multivariate experiment. In: Cattell, R. B. (Ed.): *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 459-475.
- Dipboye, R. L. & Flanagan, M. F. 1979. Research settings in industrial and organizational psychology. Are findings in the field more generalizable than in the laboratory? *American Psychologist*, 34, 141-150.
- Dixon, W. J. & Massey, F. J. Jr. 1951, 1957<sup>2</sup>, 1969<sup>3</sup>. *Introduction to statistical analysis*. New York: McGraw-Hill.
- Dixon, W. J. & Tukey, J. W. 1968. Approximate behavior of the distribution of Winsorized t (Trimming/Winsorization II). *Technometrics*, 10, 83-98.

- Dodd, D. H. & Schultz, R. F. Jr. 1973. Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin*, 79, 391-395.
- Dooling, D. J. & Danks, J. H. 1975. Going beyond tests of significance: Is psychology ready? *Bulletin of the Psychonomic Society*, 5, 15-17.
- Draper, N. R. & Hunter, W. G. 1969. Transformations: Some examples revisited. *Technometrics*, 11, 23-40.
- Draper, N. R. & Smith, H. 1966. *Applied regression Analysis*. New York: Wiley.
- Dunn, O. J. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56, 52-64.
- Dunn, O. J. & Massey, F. J. 1965. Estimation of multiple contrasts using t-distributions. *Journal of the American Statistical Association*, 60, 573-583.
- Dunnett, C. W. 1970. Multiple comparisons. In: McArthur, J. W. & Colton, Th. (Eds.): *Statistics in endocrinology*. Cambridge, Mass.: The MIT Press, 79-103.
- Dwyer, J. H. 1974. Analysis of variance and the magnitude of effects: A general Approach. *Psychological Bulletin*, 81, 731-737.
- Edgington, E. S. 1964(a). A tabulation of inferential statistics used in psychology journals. *American Psychologist*, 19, 202-203.
- Edgington, E. S. 1964(b). Randomization test. *Journal of Psychology*, 57, 445-449.
- Edgington, E. S. 1965. The assumption of homogeneity of variance for the t-test and nonparametric tests. *Journal of Psychology*, 59, 177-179.
- Edgington, E. S. 1966. Statistical inference and nonrandom samples. *Psychological Bulletin*, 66, 485-487.
- Edgington, E. S. 1969(a). Approximate randomization tests. *Journal of Psychology*, 72, 143-149.
- Edgington, E. S. 1969(b). *Statistical inference: The distribution-free approach*. New York: McGraw-Hill.
- Edgington, E. S. 1973. The random-sampling assumption in: „Comment on component-randomization tests“. *Psychological Bulletin*, 80, 84-85.
- Edgington, E. S. 1974. A new tabulation of statistical procedures used in APA journals. *American Psychologist*, 29, 25-26.
- Edwards, A. L. 1950, 1960<sup>2</sup>, 1968<sup>3</sup>, 1971. *Experimental design in psychological research*. New York: Holt, Rinehart & Winston. *Versuchsplanung in der psychologischen Forschung*. Weinheim: Beltz. Frankfurt am Main: Fachbuchhandlung für Psychologie.
- Edwards, A. L. 1957(a). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.
- Edwards, A. L. 1957(b). *The social desirability variable in personality assessment and research*. New York: Holt, Rinehart & Winston.
- Edwards, A. L. 1970. *The measurement of personality traits by scales and inventories*. New York: Holt. Rinehart & Winston.

- Edwards, A. W. F. 1972. Likelihood. Cambridge: Cambridge University Press.
- Edwards, W., Lindman, H. & Savage, L. J. 1963. Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Eimer, E. 1978. Varianzanalyse. Stuttgart: Kohlhammer.
- Eisenhart, C. 1947. The Assumptions underlying the analysis of variance. *Biometrics*, 3, 1-21.
- Einot, I. & Gabriel, K. R. 1975. A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, 70, 574-583.
- Ekbohm, G. 1976. Testing the equality of several means with small samples. *Biometrische Zeitschrift*, 18, 547-553.
- Elashoff, J. D. 1969. Analysis of covariance: A delicate instrument. *American Educational Research Journal*, 6, 383-401.
- Elashoff, R. M. 1968. Errors: Effects of errors in statistical assumptions. In: Sills, D. L. (Ed.): *International Encyclopedia of the Social sciences*. Vol. 5. Boston: Crowell Collier & Macmillan, 132-142.
- Elston, R. C. 1961. On additivity in the analysis of variance. *Biometrics*, 17, 209-219.
- Elston, R. C. & Bush, N. 1964. The hypotheses that can be tested when there are interactions in the analysis of variance model. *Biometrics*, 20, 681-698.
- Enderlein, G. (Hrsg.) (Autorenkollektiv). 1972. *Biometrische Versuchsplanung*. Berlin: VEB Deutscher Landwirtschaftsverlag.
- Engelhardt, W. 1979. Nonparametrische Prüfung von Wechselwirkungen: Verteilungsfunktionen der Interaktionsstatistik. *Psychologische Beiträge*, 21, 223-236.
- Erlebacher, A. 1977. Design and analysis of experiments contrasting the within- and between-subjects manipulation of the independent variable. *Psychological Bulletin*, 84, 212-219.
- Erlebacher, A. 1978. The analysis of multifactor experiments designed to contrast the within- and between-subjects manipulation of the independent variable. *Behavior Research Methods and Instrumentation*, 10, 833-840.
- Evans, S. H. & Anastasio, E. J. 1968. Misuse of covariance when treatment effect and covariate are confounded. *Psychological Bulletin*, 69, 225-234.
- Everitt, B. S. 1979. A Monte Carlo investigation of the robustness of Hotelling's one- and two-sample  $T^2$  tests. *Journal of the American Statistical Association*, 74, 48-51.
- Federer, W. T. 1955. *Experimental design*. New York: Macmillan
- Field, H. S. & Armenakis, A. A. 1974. On use of multiple tests of significance in psychological research. *Psychological Reports*, 35, 427-431.
- Feir-Walsh, B. E. & Toothaker, L. E. 1974. An empirical comparison of the ANOVA F-test, normal scores test and Kruskal Wallis test under Violation of assumptions. *Educational and Psychological Measurement*, 34, 789-799.
- Feldt, L. S. 1958. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, 23, 335-353.

- Feldt, L. S. 1961. The use of extreme groups to test for the presence of a relationship. *Psychometrika*, 26, 307-316.
- Feldt, L. S. 1973. What size samples for methods/materials experiments? *Journal of Educational Measurement*, 10, 221-226.
- Fennessey, J. 1968. The general linear model: A new perspective on some familiar topics. *American Journal of Sociology*, 74(1), 1-27.
- Ferguson, G. A. 1965. *Nonparametric trend analysis*. Montreal: McGill University Press.
- Festinger, L. 1978. *Theorie der kognitiven Dissonanz*. Bern: Huber.
- Festinger, L. & Carlsmith, J. M. 1959. Cognitive consequences of forced compliance. *Journal of abnormal and social Psychology*, 58, 203-210.
- Fietkau, H.-J. 1973. *Zur Methodologie des Experimentierens in der Psychologie*. Meisenheim: Hain.
- Finn, J. D. 1969. Multivariate analysis of repeated measures data. *Multivariate Behavioral Research*, 4, 391-413.
- Finn, J. D. 1974. *A general model for multivariate analysis*. New York: Holt, Rinehart & Winston.
- Fischer, G. 1974. *Einführung in die Theorie psychologischer Tests*. Bern: Huber.
- Fisher, R. A. 1925<sup>1</sup>, 1950<sup>1</sup>, 1954<sup>12</sup>. *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
- Fisher, R. A. 1928. The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society, London, A*, 121, 654-673.
- Fisher, R. A. 1935. The fiducial argument in statistical inference. *Annals of Eugenics*, 6, 391-398.
- Fisher, R. A. 1935, 1949<sup>5</sup>, 1951<sup>6</sup>, 1953. *The design of experiments*. Edinburgh: Oliver & Boyd.
- Fisher, R. A. 1956. *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd.
- Fisher, R. A. & Mackenzie, W. A. 1923. Studies in crop Variation. II. The manurial response of different potato varieties. *Journal of Agricultural Science*, 13, 311-331.
- Fisher, R. A. & Yates, F. 1938, 1963<sup>6</sup>. *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver Boyd.
- Fisz, M. 1970. *Wahrscheinlichkeitsrechnung und mathematische Statistik*. Berlin: VEB Deutscher Verlag der Wissenschaften.
- Fleishman, A. I. 1980. Confidence intervals for correlation ratios. *Educational and Psychological Measurement*, 40, 659-670.
- Fleiss, J. L. 1969. Estimating the magnitude of experimental effects. *Psychological Bulletin*, 72, 273-276.
- Fleiss, J. L. 1973. *Statistical methods for rates and proportions*. New York: Wiley.

- Fleiss, J. L. 1976. Comment on Overall and Woodward's asserted paradox concerning the measurement of change. *Psychological Bulletin*, 83, 774-775.
- Forsyth, R. A. 1978(a). A Note on „Planning an experiment in the Company of measurement error“ by Levin and Subkoviak. *Applied Psychological Measurement*, 2, 377-381.
- Forsyth, R. A. 1978(b). Some additional comments on „Planning an experiment in the Company of measurement error“. *Applied Psychological Measurement*, 2, 386-387.
- French, J. R. P. 1956, 1972<sup>8</sup>. Feldexperimente: Änderungen in der Gruppenproduktivität. In: König, R. (Hrsg.): *Beobachtung und Experiment in der Sozialforschung*. Köln: Kiepenheuer & Witsch, 259-273.
- Frey, D. 1978. Die Theorie der kognitiven Dissonanz. In: Frey, D. (Hrsg.): *Kognitive Theorien der Sozialpsychologie*. Bern: Huber, 243-292.
- Fricke, R. 1977. Möglichkeiten zur zusammenfassenden Darstellung von unabhängigen Forschungsergebnissen zur Lehrer-Schüler-Interaktion. *Zeitschrift für erziehungswissenschaftliche Forschung*, 11, 208-215.
- Friedman, H. 1968. Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, 70, 245-251.
- Friedman, H. 1969. A simplified table for the estimation of magnitude of experimental effect. *Psychonomic Science*, 14, 193-195.
- Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675-701.
- Friedman, M. 1940. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11, 86-92.
- Fromkin, H. L. & Streufert, S. 1976. Laboratory experimentation. In: Dunnette, M. D. (Ed.): *Handbook of industrial and organizational psychology*. Chicago: Rand McNally, 415-465.
- Gabriel, K. R. 1964. A procedure for testing the homogeneity of all sets of means in analysis of variance. *Biometrics*, 20, 459-477.
- Gabriel, K. R. 1969. Simultaneous tests procedures - some theory of multiple comparisons. *Annals of Mathematical Statistics*, 40, 224-250.
- Gabriel, K. R. 1978. A simple method of multiple comparisons of means. *Journal of the American Statistical Association*, 73, 724-729.
- Gabriel, R. M. & Glavin, G. B. 1978. Multivariate data analysis in empirical research: A look on the bright side. *Pavlovian Journal of Biological Science*, 13, 93-112.
- Gabriel, R. M. & Hopkins, K. D. 1974. Relative merits of MANOVA, repeated measures ANOVA, and univariate ANOVAs for research utilizing multiple criterion measures. *Journal of Special Education*, 8, 377-389.
- Gadenne, V. 1976. *Die Gültigkeit psychologischer Untersuchungen*. Stuttgart: Kohlhammer.

- Gaebelein, J. W. & Soderquist, D. R. 1978. The Utility of within-subjects variables: Estimates of Strength. *Educational and Psychological Measurement*, 38, 351-360.
- Gaebelein, J. W., Soderquist, D. R. & Powers, W. A. 1976. A note on variance explained in the mixed analysis of variance model. *Psychological Bulletin*, 83, 1110-1112.
- Gaensslen, H. & Schubö, W. 1973, 1976<sup>2</sup>. *Einfache und komplexe statistische Analyse*. München: Reinhardt.
- Gaito, J. 1958. The Bolles-Messick coefficient of Utility. *Psychological Reports*, 595-598.
- Gaito, J. 1959(a). Multiple comparisons in analysis of variance. *Psychological Bulletin*, 56, 392-393.
- Gaito, J. 1959(b). Non-parametric methods in psychological research. *Psychological Reports*, 5, 115-125.
- Gaito, J. 1960(a). Expected mean squares in analysis of variance techniques. *Psychological Reports*, 7, 3-10.
- Gaito, J. 1960(b). Scale classification and statistics. *Psychological Review*, 67, 277-278.
- Gaito, J. 1973. Repeated measurements designs and tests of null-hypothesis. *Educational and Psychological Measurement*, 33, 69-75.
- Gaito, J. 1977(a). Directional and nondirectional alternative hypotheses. *Bulletin of the psychonomic Society*, 9, 371-372.
- Gaito, J. 1977(b). Equal and unequal n and equal and unequal intervals in trend analyses. *Educational and Psychological Measurement*, 37, 283-289.
- Gaito, J. 1978. Multiple comparisons within ANOVA using orthogonal or nonorthogonal components. *Educational and Psychological Measurement*, 38, 901-904.
- Gaito, J. 1980. Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 564-567.
- Gaito, J. & Firth, J. 1973. Procedures for estimating magnitude of effects. *Journal of Psychology*, 83, 151-161.
- Gaito, J. & Turner, E. D. 1963. Error terms in trend analyses. *Psychological Bulletin*, 60, 464-474.
- Garnes, P. A. 1966. Comments on „A power comparison of the F and L tests - I“. *Psychological Review*, 73, 372-375.
- Garnes, P. A. 1971(a). Inverse relation between the risks of type I and type II errors and suggestions for the unequal n case in multiple comparisons. *Psychological Bulletin*, 97-102.
- Garnes, P. A. 1971(b). Multiple comparisons of means. *American Educational Research Journal*, 8, 531-565.
- Garnes, P. A. 1973. Type IV errors revisited. *Psychological Bulletin*, 80, 304-307.

- Garnes, P. A. 1978(a). A three-factor model encompassing many possible statistical tests on independent groups. *Psychological Bulletin*, 85, 168-182.
- Garnes, P. A. 1978(b). A four-factor structure for parametric tests on independent groups. *Psychological Bulletin*, 85, 661-672.
- Garnes, P. A. 1978(c). Nesting, crossing, type IV errors, and the role of statistical models. *American Educational Research Journal*, 15, 253-258.
- Garnes, P. A. & Howell, J. F. 1976. Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, 1, 113-125.
- Garnes, P. A., Keselman, H. K. & Clinch, J. J. 1979(a). Multiple comparisons for variance heterogeneity. *British Journal of Mathematical and Statistical Psychology*, 32, 133-142.
- Garnes, P. A., Keselman, H. K. & Clinch, J. J. 1979(b). Test for homogeneity of variance in factorial design. *Psychological Bulletin*, 86, 978-984.
- Garnes, P. A. & Lucas, P. A. 1966. Power of the analysis of variance of independent groups on non-normal and normally transformed data. *Educational and Psychological Measurement*, 26, 311-327.
- Garnes, P. A., Winkler, H. B. & Probert, D. A. 1972. Robust test for homogeneity of variance. *Educational and Psychological Measurement*, 32, 887-909.
- Gardner, P. L. 1975. Scales and statistics. *Review of Educational Research*, 45, 43-57.
- Garten, H.-K. 1980. Zur inhaltlichen Problematik von Aptitude-Treatment-Interaktion. *Zeitschrift für erziehungswissenschaftliche Forschung*, 14, 13-35.
- Gartside, P. S. 1972. A study of methods for comparing several variances. *Journal of the American Statistical Association*, 67, 342-346.
- Geary, R. C. 1936. The distribution of „Student's“ ratio for non-normal samples. *Journal of the Royal Statistical Society, Supplement*, 3, 178-189.
- Geary, R. C. 1947. Testing for normality. *Biometrika*, 34, 209-242.
- Gebert, A. 1977. Prüfung von Wechselwirkungen in  $2^2$ -faktoriellen Blockplänen mittels simultaner W-Tests. *Psychologische Beiträge*, 19, 121-129.
- Gebhardt, F. 1966. Verteilung und Signifikanzschranken des 3. und 4. Stichprobenmomentes bei normalverteilten Variablen. *Biometrische Zeitschrift*, 8, 219-241.
- Geisser, S. & Greenhouse, S. W. 1958. An extension of Box's results on the use of the F distribution in multivariate analysis. *The Annals of Mathematical Statistics*, 29, 885-891.
- Gibbons, J. D. 1971. *Nonparametric statistical inference*. New York: McGraw-Hill.
- Gibbons, J. D. & Pratt, J. W. 1975. P-Values: Interpretation and methodology. *The American Statistician*, 29(1), 20-25.
- Glass, G. V. 1966. Testing homogeneity of variances. *American Educational Research Journal*, 3, 187-190.
- Glass, G. V. 1976. Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3-8.

- Glass, G. V. 1978. Integrating findings: The meta-analysis of research. *Review of Research in Education*, 5, 351-379.
- Glass, G. V. & Hakstian, A. R. 1969. Measures of association in comparative experiments: Their developments and interpretation. *American Educational Research Journal*, 6, 40-14.
- Glass, G. V., Peckham, P. D. & Sanders, J. R. 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42, 237-288.
- Glass, G. V. & Stanley, J. C. 1970. *Statistical methods in education and psychology*. New Jersey: Prentice Hall.
- Glass, G. V., Willstott, V. L. & Gottman, J. M. 1975. *Design and analysis of time-series-experiments*. Boulder, Col.: University Press.
- Gniech, G. 1976. *Störeffekte in psychologischen Experimenten*. Stuttgart: Kohlhammer.
- Gokhale, D. V. & Kullback, S. 1978. *The information in contingency tables*. New York: Dekker.
- Gold, D. 1969. Statistical tests and Substantive significance. *The American Sociologist*, 4, 42-46.
- Golhar, M. B. 1972. The errors of first and second kinds in Welch-Aspin's Solution of the Behrens-Fisher problem. *Journal of Statistical Computation and Simulation*, 1, 209-224.
- Goodman, L. A. (Ed.) 1978. *Analyzing qualitative/categorical data*. Cambridge, Mass.: Abt Books.
- Gray, H. L. & Schucany, W. R. 1972. *The generalized jackknife statistic*. New York: Dekker.
- Greenhouse, S. W. & Geisser, S. 1959. On methods in the analysis of profile data. *Psychometrika*, 24, 95-112.
- Greenwald, A. G. 1975. Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1-20.
- Greenwald, A. G. 1976. Within-subjects designs: To use or not to use? *Psychological Bulletin*, 83, 314-320.
- Grizzle, J. E., Starmer, C. F. & Koch, G. 1969. Analysis of categorical data by linear models. *Biometrics*, 25, 489-504.
- Groeben, N. & Westmeyer, H. 1975. *Kriterien psychologischer Forschung*. München: Juventa.
- Gruijter, D. N. M. de & Kamp, L.J. T. van der (Eds). 1976. *Advances in psychological and educational measurement*. London: Wiley.
- Guenther, W. C. 1964. *Analysis of variance*. Englewood Cliffs, N. J.: Prentice Hall.
- Guenther, W. C. 1965, 1973<sup>2</sup>. *Concepts of statistical inference*. New York: McGraw-Hill. Tokyo: McGraw-Hill Kogakusha.



- Guilford, J. P. & Fruchter, B. 1956, 1973<sup>5</sup>. Fundamental statistics in psychology and education. New York: McGraw-Hill. Tokyo: McGraw-Hill.
- Guttman, L. 1977. What is not what in statistics. *The Statistician*, 26, 81-107.
- Haagen, K. & Pertler, R. 1976. Methoden der Statistik für Psychologen. Band I. Stuttgart: Kohlhammer.
- Haagen, K. & Seifert, H.-G. 1979. Methoden der Statistik für Psychologen. Band II. Stuttgart: Kohlhammer.
- Haberman, S. J. 1978. Analysis of qualitative data. Vol. 1: Introductory topics. 1978. New York: Academic Press.
- Hacking, J. 1965, 1976<sup>2</sup>. Logic of statistical inference. London: Cambridge University Press.
- Hager, W. Zur statistischen Prüfung von Interaktionshypothesen. In: Luer, G. (Hrsg.): Bericht über den 33. Kongreß der DGfPs 1982 in Mainz. Göttingen: Hogrefe (im Druck, a).
- Hager, W. Über univariate parametrische Maße der wissenschaftlichen Signifikanz. *Zeitschrift für Psychologie* (im Druck, b).
- Hager, W., Lübbecke, B. & Hübner, R. Verletzung der Annahmen bei Zwei-Stichproben-Lokationstests. *Zeitschrift für experimentelle und angewandte Psychologie* (im Druck).
- Hager, W. & Westermann, R. Zur Wahl und Prüfung statistischer Hypothesen in psychologischen Untersuchungen. *Zeitschrift für experimentelle und angewandte Psychologie* (im Druck, a).
- Hager, W. & Westermann, R. Entscheidungen über statistische und wissenschaftliche Hypothesen: Probleme bei mehrfachen Signifikanztests zur Prüfung einer wissenschaftlichen Hypothese. *Zeitschrift für Sozialpsychologie* (im Druck, b).
- Hakstian, A. R., Roed, J. C. & Lind, J. C. 1979. Two Sample  $T^2$  procedure and the assumption of homogenous covariance matrices. *Psychological Bulletin*, 86, 1255-1263.
- Halderson, J. S. & Glasnapp, D. R. 1972. Generalized rules für calculating the magnitude of an effect in factorial and repeated measures ANOVA designs. *American Educational Research Journal*, 9, 301-310.
- Hall, I. J. 1972. Some comparisons of tests for equality of variances. *Journal of Statistical Computation and Simulation*, I, 183-194.
- Hamer, R. M. & Hosking, J. D. 1977. The nonorthogonal analysis of variance. *Representative Research in Social Psychology*, 8, 71-87.
- Hamilton, B. L. 1976. A Monte Carlo test of the robustness of parametric and non-parametric analysis of covariance against unequal regression slopes. *Journal of the American Statistical Association*, 71, 864-869.
- Hamilton, B. L. 1977. An empirical investigation of the effects of heterogenous regression slopes in analysis of covariance. *Educational and Psychological Measurement*, 37, 701-712.

- Hammersley, J. M. & Handscomb, D. C. 1964. Monte Carlo methods. New York: Wiley.
- Hampel, F. R. 1980. Robuste Schätzungen: Ein anwendungsorientierter Überblick. *Biometrical Journal*, 22, 3-21.
- Harnatt, J. 1975. Der statistische Signifikanztest in kritischer Betrachtung. *Psychologische Beiträge*, 17, 595-612.
- Harnatt, J. 1979. Anmerkungen zur Kritik psychologischer Forschungsmethoden. *Psychologische Beiträge*, 21, 203-211.
- Harris, C. W. (Ed.) 1963. Problems in measuring Change. Madison: The University of Wisconsin Press.
- Harris, D. B., Bisbee, C. I. & Evans, S. H. 1971. Further comments - Misuse of analysis of covariance. *Psychological Bulletin*, 75, 220-222.
- Harris, R. J. 1975. A primer of multivariate statistics. New York: Academic Press.
- Harter, H. L. 1957. Error rates and sample sizes for range tests in multiple comparisons. *Biometrics*, 13, 511-536.
- Hartley, H. 1950. Use of range in analysis of variance. *Biometrika*, 37, 271-280.
- Havlicek, L. L. & Peterson, N. L. 1974. Robustness of the t test: A guide for researchers on effect of Violation of assumptions. *Psychological Report*, 34, 1095-1114.
- Hawkins, D. M. 1979. Fractiles of an extended multiple outlier test. *Journal of Statistical Computation and Simulation*, 8, 227-236.
- Hays, W. L. 1963, 1972, 1973<sup>2</sup>, 1977<sup>2</sup>, 1981<sup>3</sup>. Statistics for the social sciences. London: Holt, Rinehart & Winston.
- Heckhausen, H. 1969. Allgemeine Psychologie in Experimenten. Göttingen: Hogrefe.
- Hedayat, A. & Afsarinejad, K. 1975. Repeated measurements designs, I. In: Srivastava, J. N. (Ed.): A Survey of statistical design and linear models. Amsterdam: North-Holland, 229-242.
- Heermann, E. F. & Braskamp, L. A. (Eds). 1970. Readings in statistics for the behavioral sciences. Englewood Cliffs, N. J.: Prentice Hall.
- Hegemann, V. & Johnson, D. E. 1976. The power of two tests for nonadditivity. *Journal of the American Statistical Association*, 71(356), 945-948.
- Hempel, C. G. 1965. Aspects of scientific explanation. In: Hempel, C. G. (Ed.): Aspects of scientific explanation and other essays in the philosophy of science. New York: Free Press, 331-396.
- Hempel, C. G. 1974. Grundzüge der Begriffsbildung in der empirischen Wissenschaft. Düsseldorf: Bertelsmann Universitätsverlag.
- Hempel, C. G. & Oppenheim, P. 1948. Studies in the logic of explanation. *Philosophy of Science*, 15, 135-175.
- Henning, H. J. 1978. Interaktion, Transformation und empirische Bedeutsamkeit. *Archiv für Psychologie*, 130, I.: 120-138; II: 236-264.
- Henning, H. J. & Muthig, K. 1979. Grundlagen konstruktiver Versuchsplanung. München: Kösel.

- Henze, F. H.-H. 1979. The exact noncentral distributions of Spearman's  $r$  and other related correlation coefficients. *Journal of the American Statistical Association*, 74, 459-464.
- Herr, D. G. & Gaebelin, J. 1978. Nonorthogonal two-way analysis of variance. *Psychological Bulletin*, 85, 207-216.
- Herrmann, T. 1969, 1976<sup>3</sup>. *Lehrbuch der empirischen Persönlichkeitsforschung*. Göttingen: Hogrefe.
- Herrmann, T. 1973. *Persönlichkeitsmerkmale*. Stuttgart: Kohlhammer.
- Hersen, M. & Barlow, D. H. 1976. *Single-case experimental design*. New York: Pergamon.
- Higbee, K. L. & Wells, M. G. 1972. Some research trends in social psychology during the 1960s. *American Psychologist*, 27, 963-966.
- Hochberg, Y. 1976. A modification of the T-method of multiple comparisons for a oneway-layout with unequal variances. *Journal of the American Statistical Association*, 71, 200-203.
- Hochberg, Y. & Lachenbruch, P. A. 1976. Two-stage multiple comparison procedures based on the studentized range. *Communications in Statistics*, A5(15), 1447-1453.
- Hodges, J. L. & Lehmann, E. L. 1954. Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society*, B, 16, 261-268.
- Hodges, J. L. & Lehmann, E. L. 1956. The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*, 27, 324-335.
- Hogg, R. V. 1974. Adaptive robust procedures: A partial review and some suggestions for further applications and theory. *Journal of the American Statistical Association*, 69, 909-937.
- Hogg, R. V. 1977. An introduction to robust procedures. *Communications in Statistics*, A6(9), 789-794.
- Hogg, R. V. 1979. Statistical robustness: One view of its use in applications today. *The American Statistician*, 33(3), 108-115.
- Hogg, R. V. & Craig, A. T. 1970<sup>3</sup>, 1971<sup>2</sup>. *Introduction to mathematical statistics*. New York: McMillan. London: Collier McMillan International Edition.
- Hohander, M. & Wolfe, D. A. 1973. *Nonparametric statistical methods*. New York: Wiley.
- Hollingsworth, H. H. 1980. An analytical investigation of the effects of heterogeneous regression slopes in the analysis of covariance. *Educational and Psychological Measurement*, 40, 611-618.
- Holm, K. (Hrsg.) 1975, 1976, 1977. *Die Befragung*. Bde. 1-5. München: Francke.
- Holzkamp, K. 1964. *Theorie und Experiment in der Psychologie*. Berlin: de Gruyter.
- Hopkins, K. D. 1976. A simplified method for determining for expected mean squares and error terms in the analysis of variance. *Journal of Experimental Education*, 45, 13-18.

- Hopkins, K. D. & Anderson, B. L. 1973. A guide to multiple-comparison techniques: Criteria for selecting the „Method of choice“. *Journal of Special Education*, 7, 319-328.
- Horsnell, G. 1953. The effect of unequal group variances on the F-test for the homogeneity of group means. *Biometrika*, 40, 128-136.
- Hotelling, H. 1931. The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2, 360-378.
- Howell, J. F. & Garnes, P. A. 1973. The robustness of the analysis of variance and the Tukey WSD test under various patterns of heterogeneous variances. *Journal of Experimental Education*, 41, 33-37.
- Howell, J. F. & Garnes, P. A. 1974. The effects of variance heterogeneity on simultaneous multiple comparison procedures with equal sample size. *British Journal of Mathematical and Statistical Psychology*, 27, 72-81.
- Hoyle, M. H. 1973. Transformations - An introduction and a bibliography. *International Statistical Review*, 41, 203-223.
- Hsu, L. M. 1978. A Poisson method of controlling the maximum tolerable number of Type I errors. *Perceptual and Motor Skills*, 46, 211-218.
- Huber, P. J. 1972. The 1972 Wald lecture: Robust statistics: A review. *Annals of Mathematical Statistics*, 43, 1041-1067.
- Hubert, L. J. 1973. The use of orthogonal polynomials for trend analysis. *American Educational Research Journal*, 10, 241-244.
- Huberty, C. J. 1972. Multivariate indices of strength of association. *Multivariate Behavior Research*, 7, 523-526.
- Huberty, C. J. & Mourad, S. A. 1980. Estimation in multiple correlation/prediction. *Educational and Psychological Measurement*, 40, 101-112.
- Huck, S. W. & Sandler, H. M. 1973. A note on the Solomon 4-group design: Appropriate statistical analyses. *Journal of Experimental Education*, 42, 54-55.
- Huck, S. W. & Sutton, C. O. 1975. Some comments concerning the use of monotonic transformation to remove the interaction in two-factor ANOVA's. *Educational and Psychological Measurement*, 35, 789-791.
- Huitema, B. E. 1980. The analysis of covariance and alternatives. New York: Wiley.
- Hummel, T. J. 1977. Comment on „An empirical comparison of the methods of least squares and unweighted means . . .“. *International Statistical Review*, 45, 299-301.
- Hummel, T. J. & Sligo, J. R. 1971. Empirical comparison of univariate and multivariate analysis of variance procedures. *Psychological Bulletin*, 76, 49-57.
- Hummell, H. J. & Ziegler, R. (Hrsg.) 1976. *Korrelation und Kausalität*. Bd. 1-3. Stuttgart: Enke.
- Humphreys, L. G. & Fleishman, A. 1974. Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. *Journal of Educational Psychology*, 66, 464-472.

- Hurlburt, R. T. & Spiegel, D. K. 1976. Dependence of F ratios sharing a common denominator mean Square. *American Statistician*, 30, 74-78.
- Huynh, H. 1978. Some approximate tests for repeated measurement designs. *Psychometrika*, 43, 161-175.
- Huynh, H. & Feldt, L. S. 1970. Conditions under which mean square ratios in repeated measurement designs have exact F-distributions. *Journal of the American Statistical Association*, 65, 1582-1589.
- Huynh, H. & Feldt, L. S. 1976. Estimations of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, 1, 69-82.
- Huynh, H. & Mandeville, G. K. 1979. Validity conditions in repeated measures designs. *Psychological Bulletin*, 86, 969-973.
- Irle, M. 1975. *Lehrbuch der Sozialpsychologie*. Göttingen: Hogrefe.
- Jacobs, K. W. 1976. A table for the determination of experimentwise error rate ( $\alpha$ ) from independent comparisons. *Educational and Psychological Measurement*, 36, 899-903.
- James, G. S. 1951. The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, 38, 329-329.
- Jennings, E. 1967. Fixed effects analysis of variance by regression analysis. *Multivariate Behavioral Research*, 2, 95-108.
- Jennings, E. 1978. Fixed effects analysis of variance with unequal cell sizes. *Journal of Experimental Education*, 46, 42-51.
- John, J. A. & Quenouille, M. H. 1977<sup>2</sup>. *Experiments: Design and analysis*. London: Griffin.
- John, P. W. M. 1971. *Statistical design and analysis of experiments*. New York: Macmillan.
- Johnson, D. E. & Graybill, F. A. 1972. An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association*, 67, 862-868.
- Jonckheere, A. R. 1954(a). A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41, 133-145.
- Jonckheere, A. R. 1954(b). A test of significance for the relation between m rankings and k ranked categories. *British Journal of Statistical Psychology*, 7, 93-100.
- Jurs, S. G. & Glass, G. V. 1971. The effect of experimental mortality on the internal and external validity of the randomized comparative experiment. *Journal of Experimental Education*, 40, 62-66.
- Kaiser, H. F. 1960. Directional statistical decisions. *Psychological Review*, 67, 160-167.
- Katzer, J. & Sodt, J. 1973. An analysis of the use of statistical testing in communication research. *Journal of Communication*, 23, 251-265.

- Kazdin, A. E. 1980. Research design in clinical psychology. New York: Harper & Row.
- Kemp, K. E. & Conover, W. J. 1973. Robustness and power of the t-test compared with some nonparametric alternatives when sampling from a Poisson distribution. *Journal of Statistical Computation and Simulation*, 2, 293-307.
- Kempthorne, O. 1952, 1973<sup>6</sup>. Design and analysis of experiments. New York: Wiley.
- Huntington, N. Y.: Krieger.
- Kempthorne, O. 1955. The randomization theory of statistical inference. *Journal of the American Statistical Association*, 50, 946-967.
- Kendall, M. G. 1970<sup>4</sup>. Rank correlation methods. London: Griffin.
- Kendall, M. G. 1973. Time-series. London: Griffin.
- Kendall, M. G. 1975. Multivariate analysis. London: Griffin.
- Kendall, M. G. & Stuart, A. The advanced theory of statistics. Vol. I, II und III. London: Griffin, Vol. 1: 1958, 1963<sup>2</sup>, 1969<sup>3</sup>; Vol. II: 1961, 1973<sup>2</sup>; Vol. III: 1966.
- Kennedy, J. J. 1970. The eta coefficient in complex ANOVA designs. *Educational and Psychological Measurement*, 30, 885-889.
- Kenny, D. A. 1979. Correlation and causality. New York: Wiley.
- Keppel, G. 1973. Design and analysis: A researcher's handbook. Englewood Cliffs, N. Y.: Prentice-Hall.
- Keren, G. & Lewis, C. 1979. Partial omega Squared for ANOVA designs. *Educational and Psychological Measurement*, 39, 119-128.
- Kerlinger, F. N. 1964, 1973. Foundations of behavioral research. New York: Holt, Rinehart & Winston. Deutsch: Grundlagen der Sozialwissenschaften. Weinheim und Basel: Beltz, Band 1: 1975, 1978<sup>2</sup>, Band 2: 1979.
- Kerlinger, F. N. & Pedhazur, E. J. 1973. Multiple regression in behavioral research. New York: Holt, Rinehart & Winston.
- Keselman, H. J. 1975. A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared. *Canadian Psychological Review*, 16, 44-48.
- Keselman, H. J. 1976. A power investigation of the Tukey multiple comparison statistic. *Educational and Psychological Measurement*, 36, 97-104.
- Keselman, H., Garnes, P. A. & Rogan, J. C. 1979a. Protecting the Overall rate of type I errors for pairwise comparisons with an omnibus test statistic. *Psychological Bulletin*, 86, 884-888.
- Keselman, H., Garnes, P. A. & Rogan, J. C. 1979b. An addendum to: „A comparison of the modified Tukey and Scheffé methods of multiple comparisons for pairwise contrasts.“ *Journal of the American Statistical Association*, 74, 626-627.
- Keselman, H. J., Garnes, P. A. & Rogan, J. 1980. Type I and type II errors in simultaneous and two-stage multiple comparison procedures. *Psychological Bulletin*, 88, 356-358.
- Keselman, H. J., Murray, R. & Rogan, J. C. 1976. Effect of very unequal group sizes

- on Tukey's multiple comparison test. *Educational and Psychological Measurement*, 36, 263-270.
- Keselman, H. J. & Rogan, J. C. 1977. An evaluation of some non-parametric and parametric tests for multiple comparisons. *British Journal of Mathematical and Statistical Psychology*, 30, 125-133.
- Keselman, H. J. & Rogan, J. C. 1978. A comparison of the modified Tukey and Scheffé methods of multiple comparisons for pairwise contrasts. *Journal of the American Statistical Association*, 73, 47-52.
- Keselman, H. J., Rogan, J. C. & Feir-Walsh, B. J. 1977. An evaluation of some nonparametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology*, 30, 213-221.
- Keselman, H. J., Rogan, J. C., Mendoza, J. L. & Breen, L. C. 1980. Testing the validity conditions of repeated measures F-tests. *Psychological Bulletin*, 87, 479-481.
- Keselman, H. J. & Toothaker, L. E. 1973. Error rates for multiple comparison methods: Some evidence concerning the misleading conclusions of Petrinoich and Hardyck. *Psychological Bulletin*, 80, 31-32.
- Keselman, H. J. & Toothaker, L. E. 1974. Comparison of Tukey's T-method and Scheffé's S-method for various numbers of all possible differences of average contrasts under violation of assumptions. *Educational and Psychological Measurement*, 34, 511-519.
- Keselman, H. J., Toothaker, L. E. & Shooter, M. 1975. An evaluation of two unequal  $n_k$  forms of the Tukey multiple comparison statistic. *Journal of the American Statistical Association*, 70, 584-587.
- Kimball, A. W. 1951. On dependent tests of significance in the analysis of variance. *Annals of Mathematical Statistics*, 22, 600-602.
- Kirk, R. E. 1968. *Experimental design: Procedures for the behavioral sciences*. Belmont, Cal.: Brooks/Cole (Wadsworth).
- Kirk, R. E. (Ed.) 1972. *Statistical issues. A Reader for the behavioral sciences*. Belmont, Cal. : Wadsworth.
- Klauer, K. J. 1973. *Das Experiment in der pädagogischen Forschung*. Düsseldorf: Schwann.
- Kleijnen, J. P. C. 1974, 1975. *Statistical techniques in Simulation*. Part I und II. New York: Dekker.
- Kleiter, G. 1969. Krise der Signifikanztests in der Psychologie. *Jahrbuch für Psychologie, Psychotherapie und medizinische Anthropologie*, 17, 144-163.
- Kleiter, G. D. 1981. *Bayes-Statistik*. Berlin: de Gruyter.
- Knapp, T. R. 1978. Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, 85, 410-416.
- Koch, J.-J. 1976. „Guter Eindruck“ und Attitüden. Ein Beitrag zum Problem der „Verfälschbarkeit“ von Einstellungsinventaren. *Archiv für Psychologie*, 128, 135-149.

- Kohr, R. L. & Garnes, P. A. 1974. Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *Journal of Experimental Education*, 43, 61-69.
- Kohr, R. L. & Garnes, P. A. 1977. Testing complex a priori contrasts on means from independent samples. *Journal of Educational Statistics*, 2, 207-216.
- Konijn, H. S. 1973. *Statistical theory of sample survey design and analysis*. Amsterdam: North-Holland.
- Kraemer, H. C. 1974. The non-null distribution of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 69, 114-117.
- Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. 1971. *Foundations of measurement*. Vol. I: Additive and polynomial representations. New York: Academic Press.
- Kranz, H. T. 1979. *Einführung in die klassische Testtheorie*. Frankfurt am Main: Verlagsbuchhandlung für Psychologie.
- Kratochwill, T. R. 1978. *Single subject research*. New York: Academic Press.
- Krause, B. & Metzler, P. 1978. Zur Anwendung der Inferenzstatistik in der psychologischen Forschung. *Zeitschrift für Psychologie*, 186, 244-267.
- Krauth, J. 1980. Ein Vergleich der Konfigurationsfrequenzanalyse mit der Methode der log-linearen Modelle. *Zeitschrift für Sozialpsychologie*, 11, 233-247.
- Krauth, J. & Lienert, G. A. 1973. *Die Konfigurationsfrequenzanalyse (KFA)*. Freiburg: Alber.
- Kreyszig, E. 1973. *Statistische Methoden und ihre Anwendungen*. Göttingen: Vandenhoeck & Ruprecht.
- Kriz, J. 1973, 1978. *Statistik in den Sozialwissenschaften*. Reinbek: Rowohlt (rororo 1080).
- Kroll, R. M. & Chase, L. J. 1975. Communication disorders: A power analytic assessment of recent research. *Journal of Communication Disorders*, 8, 237-247.
- Krüger, H.-P. 1977. Simultane U-Tests zur exakten Prüfung von Haupt- und Wechselwirkungen an 2<sup>2</sup>-faktoriellen Versuchsplänen. *Psychologische Beiträge*, 19, 110-120.
- Kruglanski, A. W. 1975. The human subject in the psychology experiment: Fact and artifact. In: Berkowitz, L. (Ed.): *Advances in experimental social psychology*. Vol. 8. New York: Academic Press, 101-147.
- Kruskal, W. H. & Wallis, W. A. 1952. Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association*, 47, 583-621.
- Ku, H. H. & Kullback, S. 1968. Interaction in multidimensional contingency tables: An information theoretic approach. *Journal of Research of the National Bureau of Standards - Mathematical Sciences*, 72B, 159-199.
- Kühler, M. 1979. *Multivariate Analyseverfahren*. Stuttgart: Teubner.
- Küttner, M. 1976. Ein verbesserter deduktiv-nomologischer Erklärungs-begriff. *Zeitschrift für allgemeine Wissenschaftstheorie*, 7, 274-297.



- Labovitz, S. 1968. Criteria for selecting a significance level: A note on the sacredness of .05. *The American Sociologist*, 3, 220-222.
- Läuter, J. 1978. Sample size requirements for the  $T^2$ -Test of MANOVA (Tables for one-way classification). *Biometrical Journal*, 20, 389-406.
- La Forge, R. 1967. Confidence intervals or tests of significance in scientific research? *Psychological Bulletin*, 68, 446-447.
- Lakatos, J. 1970. Falsification and the methodology of scientific research programs. In: Lakatos, I. & Musgrave, A. (Eds): *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press, 91-196. Deutsch: Falsifikation und die Methodologie wissenschaftlicher Forschungsprogramme. In: Lakatos, I. & Musgrave, A. (Hrsg.): *Kritik und Erkenntnisfortschritt*. Braunschweig: Vieweg, 1974.
- Lana, R. E. 1969. Pretest sensitization. In: Rosenthal, R. L. & Rosnow, R. C. (Eds): *Artifact in behavioral research*. New York: Academic Press, 121-141.
- Lana, R. E. & Lubin, A. 1963. The effect of correlation on the repeated measures design. *Educational and Psychological Measurement*, 23, 729-739.
- Lane, D. M. & Dunlap, W. P. 1978. Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, 31, 107-112.
- Langeheine, R. 1980. *Log-lineare Modelle zur multivariaten Analyse qualitativer Daten*. München: Oldenbourg.
- Lantermann, E.-D. 1976. Zum Problem der Angemessenheit eines inferenzstatistischen Verfahrens. *Psychologische Beiträge*, 18, 99-109.
- Layard, M. W. J. 1973. Robust large-sample tests for homogeneity of variances. *Journal of the American Statistical Association*, 68, 195-198.
- Lee, E. T., Desu, M. M. & Gehan, E. A. 1975. A Monte Carlo study of the power of some two-sample tests. *Biometrika*, 62, 425-432.
- Lee, M. C. 1961. Interactions, configurations, and nonadditive models. *Educational and Psychological Measurement*, 21, 797-805.
- Lee, W. 1975. *Experimental design and analysis*. San Francisco: Freeman.
- Leeb, S. & Weinberg, S. 1977. Unreliability of difference scores: A clarification of the issues. *Psychological Reports*, 40, 931-935.
- Lehmann, E. L. 1950. Some principles of the theory of testing hypotheses. *Annals of Mathematical Statistics*, 21, 1-26.
- Lehmann, E. L. 1958. Significance level and power. *Annals of Mathematical Statistics*, 29, 1167-1176.
- Lehmann, E. L. 1968. Hypothesis testing. In: Sills, D. L. (Ed.): *International Encyclopedia of the Social Sciences*. Vol. 7. New York: Macmillan and Free Press, 40-47.
- Lehmann, E. L. 1975. *Nonparametrics: Statistical methods based on ranks*. San Francisco: Holden-Day.
- Lehmann, E. L. & Shaffer, J. P. 1977. On a fundamental theorem in multiple comparison. *Journal of the American Statistical Association*, 72, 576-578.

- Leiser, E. 1978. Einführung in die statistischen Methoden der Erkenntnisgewinnung. Köln: Pahl-Rugenstein.
- Leslie, R. I. & Brown, B. M. 1966. Use of range in testing heterogeneity of variance. *Biometrika*, 53, 221-227.
- Levene, H. 1960. Robust test for equality of variances. In: Olkin, I. (Ed.): Contributions to probability and statistics. Stanford, Cal.: Stanford University Press, 278-292.
- Levin, J. R. 1971. Some new approaches and approximations to statistical power in educational research. Unpublished Manuscript. Madison: University of Wisconsin.
- Levin, J. R. 1975. Determining sample size for planned and post hoc analysis of variance comparisons. *Journal of Educational Measurement*, 12, 99-108.
- Levin, J. R. & Marascuilo, L. A. 1972. Type IV errors and interactions. *Psychological Bulletin*, 78, 368-374.
- Levin, J. R. & Marascuilo, L. A. 1973. Type IV errors and Games. *Psychological Bulletin*, 80, 308-309.
- Levin, J. R. & Subkoviak, M. J. 1977. Planning an experiment in the company of measurement error. *Applied Psychological Measurement*, 1, 331-338.
- Levin, J. R. & Subkoviak, M. J. 1978. Correcting „Planning an experiment in the Company of measurement error“. *Applied Psychological Measurement*, 2(3), 382-385.
- Levy, K. J. 1975(a). Some multiple range tests for variances. *Educational and Psychological Measurement*, 35, 599-604.
- Levy, K. J. 1975(b). Comparing the variances of several treatments with a control. *Educational and Psychological Measurement*, 35, 793-796.
- Levy, K. J. 1975(c). An empirical comparison of several multiple range test for variances. *Journal of the American Statistical Association*, 70, 186-183.
- Levy, K. J. 1975(d). An empirical comparison of the Z-variance and Box-Scheffé tests for homogeneity of variance. *Psychometrika*, 40, 519-524.
- Levy, K. J. 1978(a). An empirical study of the tube-root test for homogeneity of variance with respect to the effects of non-normality and power. *Journal of Statistical Computation and Simulation*, 7, 71-78.
- Levy, K. J. 1978(b). Some empirical power results associated with Welch's robust analysis of variance technique. *Journal of Statistical Computation and Simulation*, 8, 43-48.
- Levy, K. J. 1978(c). An empirical comparison of the ANOVA F-test with alternatives which are more robust against heterogeneity of variance. *Journal of Statistical Computation and Simulation*, 8, 49-57.
- Levy, K. J. 1978(d). A priori contrasts under conditions of variance heterogeneity. *Journal of Experimental Education*, 47, 42-45.
- Levy, K. J. 1979(a). Nonparametric large sample many-one comparisons. *Journal of Experimental Education*, 48, 45-49.

- Levy, K. J. 1979(b). Nonparametric large-sample pairwise comparisons. *Psychological Bulletin*, 86, 371-375.
- Levy, K. J., Narula, S. C. & Abrami, P. F. 1975. An empirical comparison of the methods of least squares and unweighted means for the analysis of disproportionate cell data. *International Statistical Review*, 43, 335-338.
- Levy, P. 1967. Substantive significance of significant differences between two groups. *Psychological Bulletin*, 67, 37-40.
- Lewis, C. & Keren, G. 1977. You can't have your cake and eat it too: Some considerations of the error term. *Psychological Bulletin*, 84, 1150-1154.
- Li, C. C. 1964. *Introduction to experimental statistics*. New York: McGraw-Hill.
- Lieberman, B. (Ed.) 1971. *Contemporary problems in Statistics*. New York: Oxford University Press.
- Lienert, G. A. 1962. über die Anwendung von Variablen-Transformationen in der Psychologie. *Biometrische Zeitschrift*, 4, 145-181.
- Lienert, G. A. *Verteilungsfreie Methoden in der Biostatistik*. Bd. 1 und 2. Meisenheim am Glan: Anton Hain. Bd. 1: 1973<sup>2</sup>, Bd. II: 1978<sup>2</sup>, Tafelband: 1975.
- Lienert, G. A. & Marascuilo, L. A. 1980. Comparing treatment-induced changes for k independent samples of paired observations. *Biometrical Journal*, 22, 763-777.
- Lienert, G. A. & Raatz, U. 1971. Das Rangkorrelationsverhältnis Eta-H-Quadrat als nicht-lineares Abhängigkeitsmaß. *Biometrische Zeitschrift*, 13, 407-413.
- Lindman, H. R. 1974. *Analysis of variance in complex experimental designs*. San Francisco: Freeman.
- Lindquist, E. F. 1953. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.
- Linn, R. L. & Slinde, J. A. 1977. Significance of pre- and post-test change. *Review of Educational Research*, 47, 121-150.
- Lippman, L. G. & Taylor, C. J. 1972. Multiple comparisons in complex ANOVA designs. *Journal of General Psychology*, 86, 221-223.
- Lissitz, R. W. & Chardos, S. 1975. A study of the effect of the Violation of the assumption of independent sampling upon the type I error rate of the two-group t-test. *Educational and Psychological Measurement*, 45, 353-359.
- Loftus, G. R. 1978. On interpretation of interactions. *Memory and Cognition*, 6, 312-319.
- Lord, F. M. 1953. On the statistical treatment of football numbers. *American Psychologist*, 8, 750-751.
- Lord, F. M. 1967. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304-305.
- Lord, F. M. & Novick, M. R. 1968. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.
- Lubin, A. 1961. The interpretation of significant interaction. *Educational and Psychological Measurement*, 21, 807-817.

- Lunney, G. H. 1969. Individual size for multiple t-tests. *American Educational Research Journal*, 6, 701-703.
- MacCorquodale, K. & Meehl, P. E. 1948. On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55, 95-107.
- Mace, A. E. 1964. *Sample size determination*. New York: Holt, Rinehart & Winston.
- Mann, H. B. & Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50-60.
- Marascuilo, L. A. 1966. Large-sample multiple comparisons. *Psychological Bulletin*, 65, 280-290.
- Marascuilo, L. A. & Levin, J. R. 1970. Appropriate post hoc comparisons for interaction and nested hypothesis in analysis of variance designs: The elimination of type IV errors. *American Educational Research Journal*, 7, 397-421.
- Marascuilo, L. A. & Levin, J. R. 1976. The simultaneous investigation of interaction and nested hypotheses in two-factor analysis of variance designs. *American Educational Research Journal*, 13, 61-65.
- Marascuilo, L. A. & McSweeney, M. 1967. Nonparametric post hoc comparisons for trend. *Psychological Bulletin*, 67, 401-412.
- Marascuilo, L. A. & McSweeney, M. 1977. *Nonparametric and distribution-free methods for the social sciences*. Monterey, Cal.: Brooks/Cole.
- Marks, M. R. 1951. Two kinds of experiments distinguished in terms of statistical operations. *Psychological Review*, 58, 179-184.
- Martin, C. G. 1976. Comment on Levy's „An empirical comparison of the Z-variance and Box-Scheffé tests for homogeneity of variance“. *Psychometrika*, 41, 551-556.
- Martin, C. G. & Garnes, P. A. 1977. ANOVA tests for homogeneity of variance: nonnormality and unequal samples. *Journal of Educational Statistics*, 2, 187-206.
- Massaro, D. W. 1975. *Experimental psychology and information processing*. Chicago: Rand McNally.
- Matheson, D. W., Bruce, R. L. & Beauchamp, K. L. 1970. *Introduction to experimental psychology*. New York: Holt, Rinehart & Winston.
- Mauchly, J. W. 1940. Significance test for sphericity of a normal n-variate distribution. *Annals of Mathematical Statistics*, 11, 204-209.
- May, R. B. & Konkin, P. R. 1970. A nonparametric test of an ordered hypothesis for k independent samples. *Educational and Psychological Measurement*, 30, 251-258.
- McCall, R. B. 1970. The use of multivariate procedures in developmental psychology. In: Mussen, P. H. (Ed.): *Carmichael's manual of child psychology*. Vol. I. New York: Wiley, 1366-1377.
- McCall, R. B. & Appelbaum, M. J. 1973. Bias in the analysis of repeated-measures designs: Some alternative approaches. *Child Development*, 44, 401-415.

- McGuigan, F. J. 1979. Einführung in die experimentelle Psychologie (bearbeitet von J. M. Diehl). Frankfurt am Main: Fachbuchhandlung für Psychologie.
- McHugh, R. B. 1963. Comments on „Scales and statistics: parametric and nonparametric“. *Psychological Bulletin*, 60, 350-355.
- McNemar, Q. 1949, 1955, 1962<sup>3</sup>. *Psychological statistics*. New York: Wiley.
- McNemar, Q. 1960. At random: Sense and nonsense. *American Psychologist*, 15, 295-300.
- McSweeney, M. & Katz, B. M. 1978. Nonparametric statistics: Use and nonuse. *Perceptual and Motor Skills*, 46, 1023-1032.
- McSweeney, M. & Marascuilo, L. A. 1969. Nonparametric methods. *Review of Educational Research*, 39, 728-734.
- Meehl, P. E. 1967. Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103-115.
- Meehl, P. E. 1970. Nuisance variables and the ex-Post-facto design. In: Radner, M. & Winokur, S. (Eds): *Minnesota studies in the philosophy of science*. Vol. IV: *Analyses of theories and methods of physics and psychology*. Minneapolis: University of Minnesota Press, 373-402.
- Mehta, J. S. & Srinivasan, R. 1970. On the Behrens-Fisher problem. *Biometrika*, 57, 649-655.
- Melton, A. W. 1962. Editorial. *Journal of Experimental Psychology*, 64, 553-557.
- Mendenhall, W. 1968. *Introduction to linear models and the design and analysis of experiments*. Belmont, Cal.: Wadsworth.
- Mendoza, J. L. 1980. A significance test for multisample sphericity. *Psychometrika*, 45, 495-498.
- Mendoza, J. L., Toothaker, L. E. & Crain, B. R. 1976. Necessary and sufficient conditions for F-ratios in the LxJxK factorial design with two repeated factors. *Journal of the American Statistical Association*, 71, 992-993.
- Mendoza, J. L., Toothaker, L. E. & Nicewander, W. A. 1974. A Monte Carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures design. *Multivariate Behavioral Research*, 9, 165-178.
- Menges, G. 1968, 1972<sup>2</sup>. *Grundriß der Statistik*. Teil 1: Theorie. Opladen: Westdeutscher Verlag.
- Meredith, W. M., Fredericksen, S. H. & McLaughlin, D. 1974. Statistics and data-analysis. *Annual Review of Psychology*, 25, 453-505.
- Mertens, W. 1975. *Sozialpsychologie des Experiments*. Hamburg: Hoffmann & Campe.
- Miller, G. R. 1970. Research setting: laboratory studies. In: Emmert, P. & Brooks, W. D. (Eds): *Methods of research in communications*. New York: Houghton Mifflin, 77-104.
- Miller, J. J. 1979. Maximum likelihood estimation of variance components - A Monte Carlo study. *Journal of Statistical Computation and Simulation*, 8, 175-190.

- Miller, R. G. Jr. 1966. Simultaneous statistical inference. New York: McGraw-Hill.
- Miller, R. G. 1968. Jackknifing variances. *Annals of Mathematical Statistics*, 39, 567-582.
- Miller, R. G. 1974. The jackknife - a review. *Biometrika*, 61, 1-15.
- Miller, R. G. 1977. Developments in multiple comparison methods. *Journal of the American Statistical Association*, 72, 779-788.
- Milliken, G. A. & McDonald, L. L. 1976. Linear models and their analysis in the case of missing or incomplete data: a unifying approach. *Biometrische Zeitschrift*, 18, 381-396.
- Mood, A. M., Graybill, F. A. & Boes, D. C. 1974<sup>3</sup>. Introduction to the theory of statistics. Tokyo: McGraw-Hill.
- Moosbrugger, H. 1978. Multivariate statistische Analyseverfahren. Stuttgart: Kohlhammer.
- Morrison, D. F. 1967, 1976<sup>2</sup>. Multivariate statistical methods. New York: McGraw-Hill.
- Morrison, D. E. & Henkel, R. E. (Eds). 1970. The significance test controversy. A reader. Chicago: Aldine.
- Moses, L. E. & Oakford, R. V. 1963. Tables of random permutations. Stanford, Cal.: Stanford University Press.
- Mosteller, F. 1968. Errors: Nonsampling errors. In: Sills, D. L. (Ed.): *International Encyclopedia of the Social Sciences*. Vol. Boston: Crowell Collier and Macmillan, 113-132.
- Mosteller, F. & Bush, R. R. 1954. Selected quantitative techniques. In: Lindzey, G. (Ed.): *Handbook of social psychology*. Vol. 1. Cambridge, Mass.: Addison-Wesley, 289-334.
- Mosteller, F. & Tukey, J. W. 1968. Data analysis, including statistics. In: Lindzey, G. & Aronson, E. (Eds): *Handbook of social psychology*. Vol. 2. Reading, Mass.: Addison-Wesley, 80-203.
- Mosteller, F. & Tukey, J. W. 1977. Data analysis and regression. Reading, Mass.: Addison-Wesley.
- Myers, J. L. 1972<sup>2</sup>. Fundamentals of experimental design. Boston: Allyn & Bacon.
- Namboodiri, N. K. 1972. Experimental designs in which each subject is used repeatedly. *Psychological Bulletin*, 77, 54-64.
- Namboodiri, N. K., Carter, L. F. & Blalock, H. M. Jr. 1975. Applied multivariate analysis and experimental designs. New York: McGraw-Hill.
- Narula, S. C., Abrami, P. F. & Levy, K. J. 1976. An empirical comparison of several methods for analyzing disproportionate subclass data in a two-way classification ANOVA. *International Statistical Review*, 44, 341-348.
- Neave, H. R. & Granger, C. W. J. 1968. A Monte Carlo study comparing various two-sample-tests for differences in mean. *Technometrics*, 10, 509-522.
- Nelson, P. L. & Toothaker, L. E. 1975. An empirical study of Jonckheere's non-

- parametric test of ordered alternatives. *British Journal of Mathematical and Statistical Psychology*, 28, 167-176.
- Neyman, J. 1952<sup>2</sup>. Lectures and conferences on mathematical statistics and probability. Washington: Graduate School U. S. Department of Agriculture.
- Neyman, J. & Pearson, E. S. 1933(a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, 231, 289-337.
- Neyman, J. & Pearson, E. S. 1933(b). The testing of statistical hypotheses in relation to probabilities. *Proceedings of the Cambridge Philosophical Society*, 29, 492-510.
- Neyman, J. & Pearson, E. S. 1936. Contributions to the theory of testing statistical hypotheses. Pt. I: Unbiased critical regions of type A and type A<sub>n</sub>. *Statistical Research Memoirs*, 1-37.
- Neyman, J. & Pearson, E. S. 1938. Contribution to the theory of testing statistical hypotheses. Pt. II: Certain theorems on unbiased critical regions of type A. *Statistical Research Memoirs*, 2, 25-36.
- Neyman, J. & Pearson, E. S. 1938. Contribution to the theory of testing statistical hypotheses. Pt. III: Unbiased tests of simple statistical hypotheses specifying the values of more than one unknown Parameter. *Statistical Research Memoirs*, 2, 36-57.
- Nicewander, W. A. & Price, J. M. 1978. Dependent variable reliability and the power of significance tests. *Psychological Bulletin*, 85, 405-409.
- Nunnally, J. C. 1960. The place of statistics in psychology. *Educational and Psychological Measurement*, 20, 641-650.
- Oakes, W. F. 1972. External validity and the use of real people as subjects. *American Psychologist*, 27, 959-962.
- O'Brien, R. G. 1978. Robust techniques for testing heterogeneity of variance effects in factorial designs. *Psychometrika*, 43, 327-342.
- O'Brien, R. G. 1979. An improved ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, 74, 877-880.
- Odeh, R. E. 1972. On the power of Jonckheere's k-sample rank tests against ordered alternatives. *Biometrika*, 59, 467-471.
- Odeh, R. E. & Fox, M. 1975. Sample size choice. New York: Dekker.
- Oliver, R. L. & Berger, P. K. 1980. Advisability of pretest designs in psychological research. *Perceptual and Motor Skills*, 51, 463-471.
- Olkin, I. & Pratt, J. W. 1958. Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 29, 201-211.
- Olson, C. L. 1974. Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 69, 894-908.
- Olson, C. L. 1976. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83, 579-586.

- Olson, C. L. 1979. Practical considerations in choosing a MANOVA test statistic: A rejoinder to Stevens. *Psychological Bulletin*, 86, 1350-1352.
- O'Neill, R. & Wetherill, G. B. 1971. The present state of multiple comparison methods (with discussion). *Journal of the Royal Statistical Society, Series B*, 33, 218-241.
- Opp, K.-D. & Schmidt, P. 1976. Einführung in die Mehrvariablenanalyse. Reinbek: Rowohlt.
- Orne, M. T. 1962. On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776-783.
- Orth, D. 1974. Einführung in die Theorie des Messens. Stuttgart: Kohlhammer.
- Overall, J. E. & Dalal, S. N. 1965. Design of experiments to maximize power relative to cost. *Psychological Bulletin*, 64, 339-350.
- Overall, J. E. & Klett, C. J. 1972. Applied multivariate analysis. New York: McGraw-Hill.
- Overall, J. E. & Spiegel, D. K. 1969. Concerning least squares analysis of experimental data. *Psychological Bulletin*, 72, 311-322.
- Overall, J. E. & Woodward, J. A. 1974. A simple test for heterogeneity of variance in complex factorial designs. *Psychometrika*, 39, 311-318.
- Overall, J. E. & Woodward, J. A. 1975. Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, 82, 85-86.
- Overall, J. E. & Woodward, J. A. 1976. Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. *Psychological Bulletin*, 83, 776-777.
- Overall, J. E. & Woodward, J. A. 1977(a). Nonrandom assignment and the analysis of covariance. *Psychological Bulletin*, 84, 588-594.
- Overall, J. E. & Woodward, J. A. 1977(b). Common misconceptions concerning the analysis of covariance. *Multivariate Behavioral Research*, 12, 171-185.
- Owen, D. B. 1962. Handbook of statistical tables. Reading, Mass.: Addison-Wesley.
- Page, E. B. 1963. Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 58, 216-230.
- Patry, J. L. 1979. Feldforschung in den Sozialwissenschaften. *Zeitschrift für Klinische Psychologie und Psychotherapie*, 27, 317-335.
- Patry, J. L. (Ed.) 1982. Laborforschung/Feldforschung. Bern: Huber.
- Pearson, E. S. 1931. The analysis of variance in cases of non-normal Variation. *Biometrika*, 23, 114-133.
- Pearson, E. S. 1966. Alternative tests for heterogeneity of variance; some Monte Carlo results. *Biometrika*, 53, 229-234.
- Pearson, E. S. & Hartley, H. O. Vol. 1: 1954, 1962<sup>2</sup>; Vol. II: 1972. *Biometrika tables for statisticians*. Vol. I und II. Cambridge: Cambridge University Press.



- Pearson, E. S. & Please, N. W. 1975. Relation between the shape of population distributions and the robustness of four simple test statistics. *Biometrika*, 62, 223-241.
- Pearson, K. 1911. On a correction to be made to the correlation ratio  $\eta$ . *Biometrika*, 8, 254-256.
- Pedhazur, E. J. 1977. Coding subjects in repeated measures designs. *Psychological Bulletin*, 84, 298-305.
- Perlmutter, J. & Myers, J. L. 1973. A comparison of two procedures for testing multiple contrasts. *Psychological Bulletin*, 79, 181-184.
- Petermann, F. 1978. *Veränderungsmessung*. Stuttgart: Kohlhammer.
- Petermann, F. 1981. Möglichkeiten der Einzelfallanalyse in der Psychologie. *Psychologische Rundschau*, 32, 31-48.
- Petermann, F. & Hehl, F. J. (Hrsg.) 1979. *Einzelfallanalyse*. München: Urban & Schwarzenberg.
- Petrinovich, L. F. & Hardyck, C. D. 1969. Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. *Psychological Bulletin*, 71, 43-54.
- Pfanzagl, J. Bd. 1: 1972<sup>5</sup>, Bd. II: 1978<sup>5</sup>. *Allgemeine Methodenlehre der Statistik*. Bd. I und II. Berlin: de Gruyter.
- Pfanzagl, J. 1968. *Theory of Measurement*. Würzburg: Physica.
- Philips, L. D. 1974. *Bayesian statistics for social scientists*. New York: Crowell.
- Pillai, K. C. S. 1955. Some new test criteria in multivariate analysis. *Annals of Mathematical Statistics*, 26, 117-121.
- Pillemer, D. B. & Light, R. J. 1980. Synthesizing outcomes: How to use research evidence from many studies. *Harvard Educational Review*, 50, 176-195.
- Pitman, E. J. G. 1948. *Notes on non-parametric statistical inference*. New York: Columbia University Press.
- Plomp, T. 1974. The statistical basis for aptitude-treatment interactions research: Definitions and techniques. In: Verreck, W. A. (Ed.): *Methodological problems in research and development in higher education*. Amsterdam: Swets & Zeitlinger, 293-323.
- Poor, D. S. 1973. Analysis of variance for repeated measures designs: Two approaches. *Psychological Bulletin*, 80, 204-209.
- Popper, K. R. 1976<sup>6</sup>. *Logik der Forschung*. Tübingen: J. C. B. Mohr (Paul Siebeck).
- Popper, K. R. 1979. *Ausgangspunkte*. Hamburg: Hoffmann und Campe.
- Posten, H. O. 1978. The robustness of the two sample t-test over the Pearson System. *Journal of Statistical Computation and Simulation*, 6, 295-311.
- Pratt, J. W. 1964. Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association*, 59, 665-680.

- Preiser, S. 1977. Die experimentelle Methode. In: Strube, G. (Hrsg.): Die Psychologie des 20. Jahrhunderts. Bd. 5: Binet und die Folgen. Zürich: Kindler, 102-150.
- Puri, M. L. 1965. Some distribution-free k-sample rank tests of homogeneity against ordered alternatives. *Communications on Pure and Applied Mathematics*, 18, 51-63.
- Puri, M. L. & Sen, P. H. 1971. *Nonparametric methods in multivariate analysis*. New York: Wiley.
- Raghavarao, D. 1971. *Constructions and combinatorial problems in design of Experiments*. New York: Wiley.
- Ramsey, P. H. 1978. Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, 73, 479-485.
- Ramsey, P. H. 1980. Choosing the most powerful pairwise multiple comparison procedure in multivariate analysis of variance. *Journal of Applied Psychology*, 317-326.
- Ramseyer, G. C. & Tcheng, T.-K. 1973. The robustness of the studentized range statistic to violations of the normality and homogeneity of variance assumptions. *American Educational Research Journal*, 10, 235-240.
- Rasch, D., Enderlein, G. & Herrendörfer, G. 1973. *Biometrie*. Berlin: Deutscher Verlag der Wissenschaften.
- Rasch, D., Herrendörfer, G., Bock, J. & Busch, K. 1978. *Verfahrensbibliothek Versuchsplanung und -auswertung*. Band 1 und 2. Berlin: VEB Deutscher Landwirtschaftsverlag.
- Ray, W. S. 1966. Logic for a randomization test. *Behavioral Science*, 11, 405-406.
- Redding, W. C. 1970. Research setting: Field studies. In: Emmert, P. & Brooks, W. D. (Eds): *Methods of research in communication*. Boston: Houghton Mifflin, 105-159.
- Renn, H. 1975. *Nichtparametrische Statistik*. Stuttgart: Teubner.
- Rennert, M. 1977. Einige Anmerkungen zur Verwendung von Differenzen bei der Veränderungsmessung. *Psychologische Beiträge*, 19, 100-109.
- Restle, F. & Greeno, J. G. 1970. *Introduction to mathematical psychology*. Reading: Addison-Wesley.
- Revenstorf, D. 1979. *Zeitreihenanalyse für klinische Daten*. Weinheim: Beltz.
- Robinson, J. 1973(a). The large sample power of Permutation tests for randomization models. *Annals of Statistics*, 1, 291-296.
- Robinson, J. 1973(b). The analysis of covariance under a randomization model. *Journal of the Royal Statistical Society, Ser. B*, 35, 368-376.
- Rodger, R. S. 1967. Type II errors and their decision basis. *British Journal of Mathematical and Statistical Psychology*, 20, 187-204.
- Rodger, R. S. 1973. Confidence intervals for multiple comparisons and the misuse of the Bonferroni inequality. *British Journal of Mathematical and Statistical Psychology*, 26, 58-60.

- Rodger, R. S. 1974. Multiple contrasts, factors, error rate and power. *British Journal of Mathematical and Statistical Psychology*, 27, 178-198.
- Rodger, R. S. 1975(a). The number of non-zero, post-hoc contrasts from ANOVA and error rate. *British Journal of Mathematical and Statistical Psychology*, 28, 71-78.
- Rodger, R. S. 1975(b). Setting rejection rate for contrasts selected post hoc when some nulls are false. *British Journal of Mathematical and Statistical Psychology*, 28, 214-232.
- Rodger, R. S. 1976. Tables on Stein's non-central Parameter  $D\beta_j$ ,  $v_1$ ,  $v_2$  required to set power for numerical alternatives to  $H_0$  tested by two-stage sampling Anova. *Journal of Statistical Computation and Simulation*, 5, 1-22.
- Rodger, R. S. 1978. Two-stage sampling to set sample size for post hoc tests in ANOVA with decision-based error rates. *British Journal of Mathematical and Statistical Psychology*, 31, 153-178.
- Röhr, M. 1975. Theorie und Methodik der Kovarianzanalyse. *Probleme und Ergebnisse der Psychologie*, 52, 19-45.
- Rogan, J. C. & Keselman, H. J. 1977. Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of Variation. *American Educational Research Journal*, 14, 493-498.
- Rogan, J. C., Keselman, H. J. & Breen, L.J. 1977. Assumption violations and rates of type I error for the Tukey multiple comparison test: A review and empirical investigation via a coefficient of Variation. *Journal of Experimental Education*, 46, 20-25.
- Rogan, J. C., Keselman, H. J. & Mendoza, J. L. 1979. Analysis of repeated measures. *British Journal of Mathematical and Statistical Psychology*, 32, 269-286.
- Romaniuk, J. G., Levin, J. R. & Hubert, L.J. 1977. Hypothesis testing procedures in repeated measures designs: On the road map not taken. *Child Development*, 48, 1757-1760.
- Rosenthal, R. 1969(a). Interpersonal expectations: Effects of the experimenter's hypothesis. In: Rosenthal, R. & Rosnow, R. C. (Eds): *Artifact in behavioral research*. New York: Academic Press, 181-277.
- Rosenthal, R. 1977. Biasing effects of experimenters. *ETC.*, 34, 253-264.
- Rosenthal, R. 1979. The „file drawer problem“ and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- Rosenthal, R. 1980. On telling tails when combining results of independent studies. *Psychological Bulletin*, 88, 496-497.
- Rosenthal, R. & Gaito, J. 1963. The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, 55, 33-38.
- Rosenthal, R. & Gaito, J. 1964. Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, 15, 570.
- Rosenthal, R. & Rosnow, R. L. (Eds). 1969(a). *Artifact in behavioral research*. New York: Academic Press.

- Rosenthal, R. & Rosnow, R. C. 1969(b). The volunteer subject. In: Rosenthal, R. & Rosnow, R. C. (Eds). 1969(a): Artifact in behavioral research. New York: Academic Press, 61-118.
- Rosenthal, R. & Rubin, D. B. 1978. Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 3, 377-415.
- Rosenthal, R. & Rubin, D. B. 1979. Comparing significance levels of independent studies. *Psychological Bulletin*, 86, 1165-1168.
- Rosnow, R. & Davis, D. J. 1977. Demand characteristics and the psychological experiment. *ETC.*, 34, 301-313.
- Rotton, J. & Schönemann, P. H. 1978. Power tables for analysis of variance. *Educational and Psychological Measurement*, 38, 213-229.
- Rouanet, H. & Lépine, D. 1970. Comparison between treatments in a repeated measures design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, 23, 147-163.
- Rüppell, H. 1977. Bayes-Statistik. Eine Alternative zur klassischen Statistik. *Archiv für Psychologie*, 129, 175-186.
- Rützel, E. 1979. Bayessches Hypothesentesten und warum die Bayesianer Bias-ianer heißen sollten. *Archiv für Psychologie*, 131, 211-232.
- Rützel, E. 1980. Korrektur zu E. Rützel: Bayessches Hypothesentesten und warum die Bayesianer Bias-ianer heißen sollten. *Archiv für Psychologie*, 132, 187-188.
- Rule, S. J. 1976. A general experimentwise error rate for multiple significance tests. *Perceptual and Motor Skills*, 43, 1263-1277.
- Rulon, P. J. & Brooks, W. D. 1968. On statistical tests of group differences. In: Whitla, D. K. (Ed.): *Handbook of measurement and assessment in behavioral sciences*. Reading, Mass.: Addison-Wesley, 60-99.
- Runkel, P. J. & McGrath, J. E. 1972. *Research on human behavior*. New York: Holt, Rinehart & Winston.
- Ryan, T. A. 1959. Multiple comparisons in psychological research. *Psychological Bulletin*, 56, 26-47.
- Ryan, T. A. 1962. The experiment as the unit for computing error rates. *Psychological Bulletin*, 59, 301-305.
- Ryan, T. A. 1980. Comment on „Protecting the Overall rate of type I errors for pairwise comparisons with an omnibus test statistic“. *Psychological Bulletin*, 88, 354-355.
- Sachdeva, D. 1973. Estimating strength of relationship in multivariate analysis of variance. *Educational and Psychological Measurement*, 33, 627-631.
- Sachs, L. 1968, 1974<sup>4</sup>. *Angewandte Statistik*. Berlin: Springer.
- Samiuddin, M., Hanif, M. & Asad, H. 1978. Some comparisons of the Bartlett and cube root tests of homogeneity of variance. *Biometrika*, 65, 218-221.
- Särndal, C. E. 1974. A comparative study of association measures. *Psychometrika*, 39, 165-187.

- Saniga, E. M. & Miles, J. A. 1979. Power of some goodness-of-fit tests of normality against asymmetric stable alternatives. *Journal of the American Statistical Association*, 74, 861-865.
- Sarris, V. 1968. Nicht-parametrische Trendanalysen in der klinisch-psychologischen Forschung. *Zeitschrift für experimentelle und angewandte Psychologie*, 15, 291-316.
- Schach, S. & Schäfer, T. 1978. *Regressions- und Varianzanalyse*. Berlin: Springer.
- Scheffé, H. 1959, 1961<sup>2</sup>. *The analysis of variance*. New York: Wiley.
- Scheffé, H. 1970. Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association*, 65, 1501-1508.
- Scheffley, V. M. & Schmidt, W. H. 1978. Analysis of repeated measures data: A Simulation study. *Multivariate Behavioral Research*, 13, 347-362.
- Schlesselman, J. J. 1973. Data transformation in two-way analysis of variance. *Journal of the American Statistical Association*, 68, 369-378.
- Schmidt, F. L., Hunter, J. E. & Urry, V. W. 1976. Statistical power in criterion-related Validation studies. *Journal of Applied Psychology*, 61, 473-485.
- Schmidtke, A. & Jäger, R. 1976. Tabellen zur Überprüfung der Normalität von Schiefe und Exzeß. *Biometrische Zeitschrift*, 18, 413-418.
- Schrader, R. M. & McKean, J. W. 1977. Robust analysis of variance. *Communications in Statistics*, A6, 879-894.
- Schuler, H. 1980. *Ethische Probleme in der psychologischen Forschung*. Göttingen: Hogrefe.
- Schulman, J. L., Kupst, M. J. & Suran, B. G. 1976. The worship of „p“: Significant yet meaningless research results. *Bulletin of the Menninger Clinic*, 40, 134-143.
- Schwarz, H. 1975. *Stichprobenverfahren*. München: Oldenbourg.
- Schwarzer, R. & Steinhagen, K. (Hrsg.) 1975. *Adaptiver Unterricht. Zur Wechselwirkung von Schülermerkmalen und Unterrichtsmethoden*. München: Kösel.
- Scott, W. A. 1968. Attitude measurement. In: Lindzey, G. & Aronson, E. (Eds): *Handbook of Social Psychology*. Vol. II. Reading, Mass.: Addison-Wesley, 204-272.
- Searle, S. R. 1971(a). *Linear models*. New York: Wiley.
- Searle, S. R. 1971(b). Topics in variance component estimation. *Biometrics*, 27, 1-76.
- Selg, H. 1966, 1975<sup>4</sup>. *Einführung in die experimentelle Psychologie*. Stuttgart: Kohlhammer.
- Selg, H. & Bauer, W. 1971. *Forschungsmethoden der Psychologie*. Stuttgart: Kohlhammer.
- Shaffer, J. P. 1972. Directional statistical hypotheses and comparisons among means. *Psychological Bulletin*, 77, 195-197.
- Shaffer, J. P. 1973. Defining and testing hypotheses in multidimensional contingency tables. *Psychological Bulletin*, 79, 127-141.

- Shaffer, J. P. 1974(a). Multiple comparisons with unequal sample sizes. *Psychological Reports*, 35, 572-574.
- Shaffer, J. P. 1974(b). Bidirectional unbiased procedures. *Journal of the American Statistical Association*, 69, 437-439.
- Shaffer, J. P. & Gillo, M. W. 1974. A multivariate extension of the correlation ratio. *Educational and Psychological Measurement*, 34, 521-524.
- Shapiro, S. S., Wilk, M. B. & Chen, H. J. 1968. Comparative study of various tests for normality. *Journal of the American Statistical Association*, 63, 1343-1372.
- Sheridan, C. L. 1971. *Fundamentals of experimental psychology*. New York: Holt, Rinehart & Winston.
- Shuster, J. J. & Boyett, J. M. 1979. Nonparametric multiple comparison procedures. *Journal of the American Statistical Association*, 74, 379-382.
- Siebel, W. 1965. *Die Logik des Experimentierens in den Sozialwissenschaften*. Berlin: Duncker & Humboldt.
- Siegel, S. 1956. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill. Dt.: *Nichtparametrische statistische Methoden*. Frankfurt am Main: Fachbuchhandlung für Psychologie, 1976.
- Silverman, I. 1977. *The human subject in the psychological laboratory*. New York: Pergamon Press.
- Silverstein, A. B. 1974. Relations between analysis of variance and its nonparametric analogs. *Psychological Reports*, 34, 331-333.
- Silverstein, A. B. 1978. Critical values for nonparametric multiple comparisons. *Psychological Reports*, 43, 44-16.
- Singer, B. 1979. Distribution-free methods for non-parametric problems. A classified and selected bibliography. *British Journal of Mathematical and Statistical Psychology*, 32, 1-60.
- Skipper, J. K. Jr., Guenther, A. L. & Nass, G. 1967. The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist*, 2, 16-18.
- Smart, R. G. 1964. The importance of negative results in psychological research. *Canadian Psychologist*, 5, 225-232.
- Smart, R. G. 1966. Subject selection bias in psychological research. *Canadian Psychologist*, 7a, 115-121.
- Smith, J. E. K. 1976(a). Data transformations in analysis of variance. *Journal of Verbal Learning and Verbal Behavior*, 339-346.
- Smith, J. E. K. 1976(b). Analysis of qualitative data. *Annual Review of Psychology*, 27, 487-499.
- Smith, J. L. 1972. The eta coefficient in MANOVA. *Multivariate Behavioral Research*, 7, 361-372.
- Smith, R. A. 1971. The effect of unequal group size on Tukey's HSD procedure. *Psychometrika*, 36, 31-34.

- Snedecor, G. W. & Cochran, W. G. 1972. Statistical methods. Ames, Iowa: The Iowa State University Press.
- Soderquist, D. R. & Hussian, R. A. 1978. The Utility of Utility indices. *Bulletin of the Psychonomic Society*, 11, 136-138.
- Solomon, R. L. 1949. An extension of control group design. *Psychological Bulletin*, 46, 137-150.
- Som, R. K. 1973. A manual of sampling techniques. London: Heinemann.
- Spector, P. E. 1977. What to do with significant multivariate effects in multivariate analyses of variance? *Journal of Applied Psychology*, 62, 158-163.
- Spielman, S. 1974. The logic of tests of significance. *Philosophy of Science*, 41, 211-225.
- Spielman, S. 1978. Statistical dogma and the logic of significance testing. *Philosophy of Science*, 45, 120-135.
- Spjotvoll, E. 1974. Multiple testing in the analysis of variance. *Scandinavian Journal of Statistics*, 1, 97-114.
- Spjotvoll, E. & Stoline, M. R. 1973. An extension of the T-method of multiple comparison to include the cases with unequal sample sizes. *Journal of the American Statistical Association*, 68, 975-978.
- Sprott, D. A. 1970. Note on Evans and Anastasio on the analysis of covariance. *Psychological Bulletin*, 73, 303-306.
- Stanley, J. C. 1973. Designing psychological experiments. In: Wolman, B. B. (Ed.): *Handbook of general psychology*. Englewood Cliffs, N. J.: Prentice Hall, 90-106.
- Stavig, G. R. & Acock, A. C. 1980. Coefficients of association analogous to Pearson's  $r$  for nonparametric statistics. *Educational and Psychological Measurement*, 40, 679-685.
- Steffens, F. E. 1970. Power of bivariate studentized maximum and minimum modulus tests. *Journal of the American Statistical Association*, 65, 1639-1644.
- Steger, J. A. (Ed.) 1971. *Readings in statistics*. New York: Holt, Rinehart & Winston.
- Stegmüller, W. 1973, 1974. *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie*. Band 1: Wissenschaftliche Erklärung und Begründung, 1969b, 1974(a). Band 2: Theorie und Erfahrung. 1. Halbband: Begriffsformen, Wissenschaftssprache, empirische Signifikanz und theoretische Begriffe, 1974(b). 2. Halbband: Theoriestrukturen und Theoriedynamik, 1973(c). Band 4: Personelle und statistische Wahrscheinlichkeit. 1. Halbband: Personelle Wahrscheinlichkeit und Rationale Entscheidung, 1973(a). 2. Halbband: Statistisches Schließen. Statistische Begründung. Statistische Analyse, 1973(b). Berlin: Springer.
- Stegmüller, W. Bd. 1: 1978<sup>6</sup>, Bd. 2: 1979<sup>6</sup>(a). *Hauptströmungen der Gegenwartsphilosophie*. Bd. 1 und 2. Stuttgart: Kröner.
- Stegmüller, W. 1979(b). *The structuralist view of theories*. Berlin: Springer.
- Stegmüller, W. 1980. *Neue Wege der Wissenschaftsphilosophie*. Berlin: Springer.

- Stein, C. 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16, 243-258.
- Steinfatt, T. M. 1977. Measurement, transformations, and the real world: Do the numbers represent the concept? *ETC.*, 34, 277-289.
- Sterling, T. D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, 54, 30-34.
- Sterling, T. D. 1960. What is so peculiar about accepting the null-hypothesis? *Psychological Reports*, 7, 363-364.
- Stevens, J. P. 1972(a). Global measures of association in multivariate analysis of variance. *Multivariate Behavioral Research*, 7, 373-378.
- Stevens, J. P. 1972(b). Four methods of analyzing between Variation for the k group Manova Problem. *Multivariate Behavioral Research*, 7, 499-501.
- Stevens, J. P. 1979. Comment on Olson: Choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 86, 355-360.
- Stevens, J. P. 1980. Power of the multivariate analysis of variance tests. *Psychological Bulletin*, 88, 728-737.
- Stevens, S. S. 1951. Mathematics, measurement, and psychophysics. In: Stevens, S. S. (Ed.): *Handbook of experimental psychology*. New York: Wiley, 1-49.
- Steyer, R. 1977, 1979. Untersuchungen zur nonorthogonalen Varianzanalyse. Göttingen: Diplomarbeit. Weinheim: Beltz.
- Stilson, D. W. 1966. Probability and statistics in psychological research and theory. San Francisco: Holden-Ray.
- Student (W. S. Gosset). 1908. The probable error of the mean. *Biometrika*, 6, 1-25.
- Subkoviak, M. J. & Levin, J. R. 1977. Fallibility of measurement and the power of a statistical test. *Journal of Educational Psychology*, 14, 47-52.
- Summers, G. F. (Ed.) 1970. Attitude measurement. Chicago: Rand McNally.
- Suppe, F. 1974, 1977<sup>2</sup>(a). The search for philosophic understanding of scientific theories. In: Suppe, F. (Ed.): *The structure of scientific theories*. Urbana, Ill.: University of Illinois Press, 3-232.
- Suppe, F. 1977<sup>2</sup>(b). Afterword - 1977. In: Suppe, F. (Ed.): *The structure of scientific theories*. Urbana, Ill.: University of Illinois Press, 617-630.
- Suppes, P. 1970. A probabilistic theory of causality. Amsterdam: North-Holland.
- Suppes, P. & Zinnes, J. L. 1963. Basic measurement theory. In: Luce, R. D., Bush, R. R. & Galanter, E. (Eds): *Handbook of mathematical psychology*. Vol. 1. New York: Wiley, 1-76.
- Sutcliffe, J. P. 1958. Error of measurement and the sensitivity of a test of significance. *Psychometrika*, 23, 9-17.
- Sutcliffe, J. P. 1980. On the relationship of reliability to statistical power. *Psychological Bulletin*, 88, 509-515.



- Swaminathan, H. & DeFriesse, F. 1979. Detecting significant contrasts in analysis of variance. *Educational and Psychological Measurement*, 39, 39-44.
- Talwar, P. P. & Gentle, J. E. 1977. A robust test for the homogeneity of scales. *Communications in Statistics*, A6, 363-369.
- Tamhane, A. C. 1977. Multiple comparisons in model I one-way ANOVA with unequal variances. *Communications in Statistics*, A6, 15-32.
- Tamhane, A. C. 1979. A comparison of procedures for multiple comparisons of means with unequal variances. *Journal of the American Statistical Association*, 74, 471-480.
- Tang, P. C. 1938. The power function of the analysis of variance tests with tables and illustrations of their use. *Statistical Research Memoirs*, 2, 126-149.
- Tatsuoka, M. 1969. Multivariate analysis. *Review of Educational Research*, 39, 739-743.
- Tatsuoka, M. M. 1971. *Multivariate analysis*. New York: Wiley.
- Thomas, D. A. H. 1973. Multiple comparisons among means. A review. *The Statistician*, 22, 16-42.
- Tiku, M. L. 1971. Power function of the F-test under non-normal situations. *Journal of the American Statistical Association*, 66, 913-916.
- Tiku, M. L. 1975. A new statistic for testing suspected outliers. *Communications in Statistics*, 4, 737-752.
- Timaeus, E. 1974. *Experiment und Psychologie. Zur Sozialpsychologie psychologischen Experimentierens*. Göttingen: Hogrefe.
- Timaeus, E. 1975. Untersuchungen im Laboratorium. In: Koolwijk, J. v. & Wicken-Mayser, M. (Hrsg.): *Techniken der empirischen Sozialforschung*. Bd. 2: Untersuchungsformen. München: Oldenbourg, 195-229.
- Toothaker, L. E. 1971. The effect of the joint Violation of two assumptions on the size of the F-test under Permutation for the randomized block design. *Journal of Statistical Computation and Simulation*, 1, 55-64.
- Toothaker, L. E. 1972. An empirical investigation of the Permutation t-test. *British Journal of Mathematical and Statistical Psychology*, 25, 83-94.
- Trachtman, J. N., Giambalvo, V. & Dippner, R. S. 1978. On the assumptions concerning the assumptions of a t test. *The Journal of General Psychology*, 99, 107-116.
- Traxel, W. 1974. *Grundlagen und Methoden der Psychologie*. Bern: Huber.
- Treiber, B. 1977. Untersuchungen zur Interaktion von Lehrmethode und Schülermerkmale: Reanalyse der Teststärke. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 90, 29-35.
- Treiber, B. & Petermann, F. 1976. Zur Interaktion von Lernermerkmalen und Lehrmethoden: Rekonstruktion und Normierung des ATI-Forschungsprogramms. Heidelberg: Bericht No. 4, aus dem Institut für Psychologie.
- Treinius, G. 1977. Teststärkeanalysen für 61 Untersuchungen zum Zusammenhang

- zwischen Lehrerverhalten und Schülerleistung. Zeitschrift für erziehungswissenschaftliche Forschung, 11, 50-63.
- Trickett, W. H. & Welch, B. L. 1954. On the comparison of two means, further discussion of iterative methods for calculation tables. *Biometrika*, 41, 361-374.
- Trickett, W. H., Welch, B. L. & James, G. S. 1956. Further critical values for the two-means problem. *Biometrika*, 43, 203-205.
- Tukey, J. W. 1949. One degree of freedom for nonadditivity. *Biometrics*, 5, 232-242.
- Tukey, J. W. 1957. On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, 28, 602-632.
- Tukey, J. W. 1977. *Exploratory data analysis*. Reading, Mass.: Addison-Wesley.
- Tukey, J. W. & McLaughlin, D. H. 1963. Less vulnerable confidence and significance procedures for location based on a single sample Trimming/Winsorization. 1. *Sankhyā*, A25, 331-352.
- Underwood, B. J. 1957. *Psychological research*. New York: Appleton-Century-Crofts.
- Underwood, B. J. 1966<sup>2</sup>. *Experimental psychology*. New York: Appleton-Century-Crofts.
- Underwood, B. J. & Shaughnessy, J. J. 1975. *Experimentation in psychology*. New York: Wiley.
- Upton, G. J. G. 1978. *The analysis of cross-tabulated data*. Chichester: Wiley.
- Ury, H. K. 1976. A comparison of four procedures for multiple comparisons among means pairwise contrasts for arbitrary sample sizes. *Technometrics*, 18, 89-97.
- Ury, H. K. & Wiggins, A. D. 1971. Large sample and other multiple comparisons among means. *British Journal of Mathematical and Statistical Psychology*, 24, 174-194.
- Ury, H. K. & Wiggins, A. D. 1974. Use of the Bonferroni inequality for multiple comparisons among means with post hoc contrasts. *British Journal of Mathematical and Statistical Psychology*, 27, 176-178.
- Ury, K. H. & Wiggins, A. D. 1975. A comparison of three procedures for multiple comparisons among means. *British Journal of Mathematical and Statistical Psychology*, 28, 88-102.
- Vatza, E. J., Byatt, S. E., Kay, K. J., Kerchner, M., Richter, M. L. & Seay, M. B. 1980. Comment on „Combining results of independent studies“. *Psychological Bulletin*, 88, 494-495.
- Vaughan, G. M. & Corballis, M. C. 1969. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, 72, 204-213.
- Venables, W. 1975. Calculation of confidence intervals for noncentrality Parameters. *Journal of the Royal Statistical Society, Series B*, 37, 406-412.
- Verreck, W. A. (Ed.) 1974. *Methodological problems in research and development in higher education*. Amsterdam: Swets & Zeitlinger.

- Wainer, H. 1972. A practical note on one-tailed tests. *American Psychologists*, 27, 775-776.
- Wainer, H. 1973. The other tail. *British Journal of Mathematical and Statistical Psychology*, 26, 182-187.
- Wainer, H. 1976. Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics*, 1, 285-312.
- Wainer, H. & Thissen, D. 1976. Three steps toward robust regression. *Psychometrika*, 41, 9-34.
- Wainer, H. & Thissen, D. 1981. Graphical data analysis. *Annual Review of Psychology*, 32, 191-241.
- Wald, A. 1947, 1952<sup>3</sup>. *Sequential analysis*. New York: Wiley.
- Wald, A. 1950. *Statistical decision functions*. New York: Wiley.
- Wallenstein, S. & Fleiss, J. L. 1979. Repeated measurements analysis of variance when the correlations have a certain pattern. *Psychometrika*, 44, 229-233.
- Walsh, J. E. 1962, 1965, 1968. *Handbook of nonparametric statistics*. Vol. 1: 1962, Vol. II: 1965, Vol. III: 1968. Princeton, N. Y.: Van Nostrand.
- Wang, Y. Y. 1967. A comparison of several variance component estimators. *Biometrika*, 54, 301-305.
- Wang, Y. Y. 1971. Probabilities of the Type I errors of the Welch test for the Behrens-Fisher problem. *Journal of the American Statistical Association*, 66, 605-608.
- Webb, E. J., Campbell, D. T., Schwartz, R. D. & Sechrest, L. 1966. *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally. Dt.: *Nichtreaktive Meßverfahren*. Weinheim: Beltz, 1975.
- Weber, E. 1967<sup>6</sup>. *Grundriß der biologischen Statistik*. Stuttgart: Gustav Fischer Verlag.
- Weber, S. J. & Cook, T. D. 1972. Subjects effects in laboratory research: An examination on subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, 77, 273-295.
- Wegman, E. J. & Carroll, R. J. 1977. A Monte Carlo study of robust estimators of location. *Communications in Statistics*, A6, 795-812.
- Welch, B. L. 1947. The generalization of „Student's“ problem when several populations are involved. *Biometrika*, 34, 28-35.
- Welch, B. L. 1949. Further note on Mrs. Aspin's tables and on certain approximations to the tabled function. *Biometrika*, 36, 293-296.
- Welch, B. L. 1951. On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330-336.
- Werts, C. E. & Linn, R. L. 1971. Problems with inferring treatment effects from repeated measures. *Educational and Psychological Measurement*, 31, 857-866.
- Westermann, R. 1980. Die empirische Überprüfung des Niveaus psychologischer Skalen. *Zeitschrift für Psychologie*, 188, 450-468.

- Westermann, R. 1982. Zur Messung von Einstellungen auf Intervallskalenniveau. *Zeitschrift für Sozialpsychologie*, 13, 97-108.
- Westermann, R. Zur Wahl der Fehlerwahrscheinlichkeiten bei mehrfachen Signifikanztests. In: Lüer, G. (Hrsg.): Bericht über den 33. Kongreß der Deutschen Gesellschaft für Psychologie in Mainz 1982. Göttingen: Hogrefe (im Druck, a).
- Westermann, R. Zur empirischen Überprüfung des Skalenniveaus von individuellen Einschätzungen und Ratings. *Zeitschrift für Psychologie* (im Druck, b).
- Westermann, R. & Hager, W. 1982. Entscheidung über statistische und wissenschaftliche Hypothesen: Zur Differenzierung und Systematisierung der Beziehungen. *Zeitschrift für Sozialpsychologie*, 13, 13-21.
- Westmeyer, H. 1982. Wissenschaftstheoretische Aspekte der Feldforschung. In: Patry, J. L. (Hrsg.): Laborforschung/Feldforschung. Bern: Huber, 67-84.
- Wetherill, G. B. 1960. The Wilcoxon test and non-null hypotheses. *Journal of the Royal Statistical Society, Series B*, 22, 402-418.
- Wetherill, G. B. 1966, 1975<sup>2</sup>. *Sequential methods in statistics*. London: Chapman and Hall.
- Wetzel, W. (Hrsg.) 1970. Neue Entwicklungen auf dem Gebiet der Zeitreihenanalyse. Göttingen: Vandenhoeck & Ruprecht.
- Whitney, D. R. & Feldt, L. S. 1973. Analyzing questionnaire results: Multiple tests of hypotheses and multivariate hypotheses. *Educational and Psychological Measurement*, 33, 365-380.
- Wike, E. L. & Church, J. D. 1977. Further comments on nonparametric multiple-comparison tests. *Perceptual and Motor Skills*, 45, 917-918.
- Wike, E. L. & Church, J. D. 1978. A Monte Carlo investigation of four nonparametric multiple-comparison tests for k independent groups. *Bulletin of the Psychonomic Society*, 11, 25-28.
- Wildt, A. R. & Ahtola, O. T. 1976. *Analysis of covariance*. Beverly Hills: Sage.
- Wilk, M. B. & Kempthorne, O. 1955. Fixed, mixed and random models. *Journal of the American Statistical Association*, 50, 1144-1167.
- Wilk, M. B. & Kempthorne, O. 1957. Nonadditives in a Latin square design. *Journal of the American Statistical Association*, 52, 218-236.
- Williams, J. D. 1974. A simplified regression formulation of Tukey's test. *Journal of Experimental Education*, 42, 4, 80-82.
- Wilson, K. 1975. The sampling distributions of conventional conservative and corrected F-ratios in repeated measurements designs with heterogeneity of covariance. *Journal of Statistical Computation and Simulation*, 3, 201-215.
- Wilson, R. S. 1975. Analysis of developmental data: Comparison among alternative models. *Developmental Psychology*, 11, 676-680.
- Winer, B. J. 1962, 1970, 1971<sup>2</sup>. *Statistical principles in experimental design*. Tokyo: McGraw-Hill Kogakuska.

- Winne, D. 1968. Zur Planung von Versuchen: Wieviel Versuchseinheiten? *Arzneimittelforschung*, 18, 1611-1618.
- Wishart, J. 1932. A note on the distribution of the correlation ratio. *Biometrika*, 24, 441-456.
- Witte, E. H. 1977. Zur Logik und Anwendung der Inferenzstatistik. *Psychologische Beiträge*, 19, 290-303.
- Witte, E. H. 1978. Korrelationsstudien und Experimente: Ihre Kombination und ihre Aussagekraft. Hamburg: Arbeiten aus den Psychologischen Instituten der Universität Hamburg, Nr. 46.
- Witte, E. H. 1980. Signifikanztest und statistische Inferenz. Stuttgart: Enke.
- Wooding, W. M. 1969. The computation and use of residuals in the analysis of experimental data. *Journal of Quality Technology*, 1, 175-188, 299.
- Woodward, A. J. & Overall, J. E. 1975. Multivariate Analysis of variance by multiple regression methods. *Psychological Bulletin*, 82, 21-32.
- Woodward, J. A. & Overall, J. E. 1976. A Computer program for calculating power of the F-test. *Educational and Psychological Measurement*, 36, 165-168.
- Wormser, R. 1974. *Experimentelle Psychologie*. München: Reinhardt.
- Wottawa, H. 1974. Das „allgemeine lineare Modell“ - ein universelles Auswertungsverfahren. *EDV in Medizin und Biologie*, 3, 65-73.
- Wottawa, H. 1980. *Grundriß der Testtheorie*. München: Juventa.
- Wundt, W. 1896, 1913<sup>11</sup>. *Grundriß der Psychologie*. Leipzig: Engelmann.
- Yates, F. 1933. The principles of orthogonality and confounding in replicated experiments. *Journal of Agricultural Science*, 23, 108-145.
- Young, R. K. & Veldman, D. J. 1963. Heterogeneity and skewness in analysis of variance. *Perceptual and Motor Skills*, 16, 588.
- Zimmermann, E. 1972. *Das Experiment in den Sozialwissenschaften*. Stuttgart: Teubner.
- Zimny, G. H. 1961. *Method in experimental psychology*. New York: Ronald Press.

### 3. Kapitel

## Messung, Analyse und Prognose von Veränderungen

*Claus Möbus und Willi Nagl*

### *1. Einleitung*

Psychodiagnostik läßt sich nach Pawlik (1976) auf vier Dimensionen mit entsprechend unterschiedlichen Zielsetzungen betreiben: (1) Status- vs. Prozeßdiagnostik, (2) norm- vs. kriterienorientierte Diagnostik, (3) Testen vs. Inventarisieren und (4) Diagnostik als Messung vs. Diagnostik als Information für und über Behandlung. Setzt man sich die Prozeßdiagnostik und Behandlungsoptimierung zum Ziel, muß man, wenn man empirisch arbeitet, Veränderungsmessung betreiben. Inhaltliche Beispiele für Veränderungsmessung lassen sich für alle Bereiche der Psychologie, die an „Intervention“, „Entwicklung“, „Dynamik“ etc. interessiert sind, finden (so z.B. in einigen neueren deutschsprachigen Arbeiten: Renn, 1973; Straka, 1974; Krapp & Schiefele, 1976; Kleiter & Petermann, 1977; Möbus & Wallasch, 1977; Rudinger & Lantermann, 1978; Baltes, 1979; Baltes & Nesselroade, 1979; Kormann, 1979; Metz-Göckel, 1979; Petermann & Hehl, 1979; Kleiter, 1979; Revenstorf, 1979; Kormann, 1981; Möbus, 1983). Hierunter fällt u.a. die Erfassung von „Lernprozessen“, „Lernwegen“, „Entwicklungsverläufen“, „Wachstumskurven“, „Einstellungsänderungen“ und „Therapieeffekten“.

Entsprechend der Vielfalt der inhaltlichen Fragestellungen hat sich eine kaum zu überschauende Menge verschiedener Methoden und Schulen herausgebildet, die z.T. miteinander im Wettstreit stehen (s. a. Rost & Spada, 1978). Dabei hat es sich für die Analyse von *Prozessen* als hemmend (wenn nicht gar irreführend) erwiesen, Methoden aus dem statischen Querschnitts- in den dynamischen Zeitbereich zu übertragen. Wenig brauchbar im Zusammenhang mit Veränderungsmessung scheinen uns Verfahren zu sein, die einerseits keine *Hypothesenprüfung* zulassen, ob eine Veränderung bei einem Individuum oder einer Gruppe von Individuen signifikant war und die andererseits keine *falsifizierbaren Prognosen für konkrete (apriori bestimmbare) Zeitpunkte* zulassen. Roskam vertritt eine ähnlich vorsichtige Haltung, wenn er schreibt:

„.... Perhaps the time is simply not yet ripe for doing thorough research on change processes in behaviour, and should we spend our time more wisely in doing the lot of ground work which has to be done first so that we are better equipped before starting on so much complicated problems. This is not meant to discourage research on change. *Rather it is an advice to go about vey carefully and vey cautiourly.*“ (Roskam, 1976, S. 132f.)

Akzeptiert man die beiden oben genannten Kriterien als Gütemaßstäbe für Methoden zur Analyse von Veränderungen, kann man in dem Klassiker von Harris (1963) aus heutiger Sicht weiterführende und weniger nützliche Artikel separieren. Von den 12 Beiträgen über:

(1) „Some persisting Dilemmas in the Measurement of Change“ (Bereiter), (2) „Elementary Models for Measuring Change“ (Lord), (3) „The Reliability of Changes Measured by Mental Test Scores“ (Webster & Bereiter), (4) „Univariate Analysis of Variance Procedures in the Measurement of Change“ (Gaito & Wiley), (5) „Multivariate Analysis of Variance of Repeated Measurements“ (Bock), (6) „Multivariate Models for Evaluating Change“ (Horst), (7) „Implications of Factor Analysis of Three-Way Matrices for Measurement of Change“ (Tucker), (8) „Canonical Factor Models for the Description of Change“, (Harris), (9) „Image Analysis“ (Kaiser), (10) „The Structuring of Change by P-Technique and Incremental R-Technique“ (Cattell), (11) „Statistical Models for the Study of Change in the Single Case“ (Holtzman) und (12) „From Description to Experimentation: Interpreting Trends as Quasi-Experiments“ (Campbell)

erscheinen uns die Ansätze von Bock, Holtzman und Campbell auch heute wertvoll, während die Artikel von Bereiter und Webster & Bereiter relativ viel Verwirrung über den Nutzen von Veränderungsmaßen gebracht haben. Während Bereiter Mittelwertsdifferenzen nicht in Frage stellte, formulierte er eine Reihe von Dilemmata, die sich einstellen sollten, wenn man *individuelle* Veränderungen durch die Differenz zeitlich verschobener Messungen operationalisiert: „over - correction - under - correction dilemma“, „unreliability - invalidity - dilemma“ und das „physicalism-subjectivism dilemma“. Das erste Dilemma ergibt sich nach Bereiter aus der Erkenntnis „.... that there is a spurious negative element in the correlation of an initial score with gains on the same test . . .“ (Bereiter, 1963, S. 3). Das zweite Dilemma entsteht nach Bereiter „.... that, other things being equal, the higher the correlation between pretest and post-test, the lower the reliability of the difference scores. Accounting for the other horn of the dilemma is the even more elementary fact that the lower the correlation between two tests, the less they can be said to measure the same thing“ (Bereiter, 1963, S. 5). Das dritte Dilemma bestand in der Frage, ob die gleiche Veränderung in der manifesten, beobachtbaren Variablen einhergeht mit gleichen Veränderungen auf der latenten Variablen, wenn die Ausgangswerte unterschiedlich hoch sind. Diese Dilemmata zogen sich wie ein roter Faden durch die Literatur und bestimmten die Diskussion für eine Reihe von Jahren.

Wir sind der Ansicht, daß es sich um Scheinprobleme handelt, wenn man einmal das Reliabilitätskonzept der klassischen Testtheorie nicht kritiklos akzeptiert und zum anderen Veränderungsanalyse unter einer systemtheoretischen Perspektive betreibt. Eine systemtheoretische Sicht der Veränderungsanalyse bedeutet: (1) der Veränderungsprozeß einer Person oder (aggregiert) einer Gruppe läßt sich kontinuierlich in der Zeit multivariat (probabilistisch oder nichtprobabilistisch) beschreiben. Die endogene Entwicklung wird überlagert durch Effekte exogener Variabler (äußere Rahmenbedingungen), die nur unvollständig kontrolliert werden können. Dieser stochastische zeitkontinuierliche Prozeß kann nur zu bestimmten diskreten Zeitpunkten gemessen werden (Sampling). Der Meßprozeß wird durch Meßfehler gestört. Selbst wenn die Prozeßparameter für 2 Personen gleich sind, können die Entwicklungsverläufe aufgrund unterschiedlicher Ausgangsbedingungen (Anfangszustände) und/oder unterschiedlicher exogener Rahmenbedingungen für die 2 Personen differieren (nähere Erläuterung im Abschnitt über zeitkontinuierliche Modelle). Eine Prognose (z.B. im Sinne von Erwartungswerten) ist nur möglich, wenn (a) der Anfangszustand (Pretest) gemessen oder geschätzt (= „wahrer Anfangszustand“) wird, (b) wenn die Gesetzhypothese (niedergelegt in den Parametern des Systems) sich nicht unvorhergesehen verändert und (c) wenn die exogenen Einflüsse (Rahmenbedingungen) entweder kontrolliert oder prognostiziert werden können oder konstant bleiben. Prognosen beziehen sich immer nur auf bestimmte konkrete Zeitpunkte. Ebenso sind Hypothesen auch nur an bestimmte Zeitpunkte geknüpft.

In dieser Sicht ist es völlig natürlich, interindividuell unterschiedliche Entwicklungskurven zu erwarten, wenn nur die äußeren Einflüsse und/oder die Ausgangsbedingungen für Individuen über die Zeit verschieden sind. Es ist nicht notwendig (wie Bereiter) zu glauben, niedrige Pre-Posttest-Korrelationen müßten unbedingt auf eine veränderte Bedeutung der Meßinstrumente verweisen. Sie kann vielmehr ein Hinweis auf zwischenzeitlich wirkende äußere Einflüsse sein. Es ist auch nicht sinnvoll, bei dem Studium von Methoden der Veränderungsanalyse mit Korrelationen zu argumentieren, wie es z.B. in der Reliabilitätsdiskussion getan wird, denn bei Korrelationen gehen Mittelwerte und Streuungen verloren. Was bei heterogenen Variablen mit unterschiedlichen Maßstäben in Querschnittanalysen sinnvoll sein kann, ist bei Veränderungsmessungen verfehlt, es sei denn, wir trauten selbst Mittelwerts- und Streuungsveränderungen keinen Informationsgehalt mehr zu. Dann wäre Veränderungsmessung aber weitgehend sinnlos. Wichtig sind sauber operationalisierte Meßvariable, deren praktische Bedeutsamkeit auf der Hand liegt und nicht im Nachhinein durch interpretative Vorgänge erst erschlossen werden muß. Eine ähnliche Ansicht vertritt z.B. Roskam und Rost & Spada, wenn sie schreiben:

„Researchers often believe that it is wrong to analyse change or gain scores: difference scores are unreliable, and if they are very reliable, how do we know whether it is the



persons or the tests who have changed (cf. Bereiter, 1963)? If a pretest and a posttest correlate as high as their reliability, the reliability of the difference scores is zero, but if they correlate appreciably less than their reliability, one is led to believe that they measure different constructs, and the difference scores are hard to interpret. Bereiter quotes Jordan in saying: „it is senseless to consider a test a valid measure of an attribute that is not clearly conceptualized independently of any instruments supposed to measure it“. So without conceptualizing first what we intend to measure, it would be meaningless to talk about change. It is my firm belief that it is methodological nonsense. Test scores - or any other kind of controlled observation - are factual data; if a persons reactions on a second occasion are different from those on a first occasion, it is our task to find a (theoretical) explanation for that fact.“ (Roskam, 1979, S. 2).

„Findet jedoch zwischen beiden Zeitpunkten, zu denen die Tests vorgegeben wurden, personenspezifisches Lernen statt, d.h. ändern sich die wahren Werte der Versuchspersonen zwischen den Testzeitpunkten, dann ist es sinnlos, anhand der Korrelation der Tests etwas über ihre Äquivalenz auszusagen. Beide Tests sollen ja in diesem Fall gar keine äquivalenten Messungen liefern, sondern einmal die Fähigkeit zum Zeitpunkt  $t_1$ , das andere Mal die zum Zeitpunkt  $t_2$  messen.“ (Rost & Spada, 1978, S. 90f.)

Wir wollen hier davon ausgehen, daß die Variablen, auf denen Veränderungen gemessen werden, eine Bedeutung haben, die nicht erst nach den Messungen durch korrelative Studien interpretativ „festgelegt“ werden muß.

Auch das Problem nichtsignifikanter Interventionen erfährt unter systemtheoretischer Beleuchtung eine gewisse Aufhellung. So ist man, wenn enge Rahmenbedingungen hinsichtlich der zur Verfügung stehenden Ressourcen vorgegeben sind, gezwungen, Entscheidungs- bzw. Behandlungsoptimierung zu betreiben. In den meisten Fällen läßt sich die zeitliche Dauer der Intervention (Therapie, Training, Instruktion etc.) und der den Prozeßverlauf abbildenden Messungen aus verschiedenen Gründen (Ermüdung der Pbn, Kosten der Datenerhebung etc.) nicht beliebig verlängern. Daher stellt sich oft die Aufgabe, bei fester Test- und Interventionsdauer die „Mixtur“ der Trainingsteile optimal zu verändern. Läßt sich der Interventionseffekt nicht statistisch sichern, wird damit meist vorschnell auf die mangelnde *Güte* der Intervention geschlossen. Eine genaue Analyse zeigt aber einen weitaus komplizierteren Sachverhalt. Es gibt mindestens drei Erklärungsmöglichkeiten für die Nichtsignifikanz: (a) mangelnde Güte des Interventionsprogramms, (b) zu kurze zeitliche „*Dosierung*“ eines an sich effektiven Programms, (c) Flüchtigkeit des Effekts: nach dem Absetzen der Intervention verschwindet der Effekt (liegen die Meßzeitpunkte weit auseinander, wird oft gar kein Effekt wahrgenommen). Die letzten beiden Erklärungen sowie die Planung der Optimierung können mit klassischen statistischen Verfahren (Varianz-, Faktorenanalyse etc.) nur unvollständig geleistet werden. Eine Untersuchung zum Dosierungsproblem eines Trainings und der Versuch einer Kreuzvalidierung an einer Längsschnittstudie findet sich bei Möbus (1981). Verwendet wurden dabei stochastische zeitdiskrete Differenzengleichungssysteme, deren Parameter mit LISREL geschätzt wurden (s.a. 7.3).

Unter dem oben angerissenen Blickwinkel haben wir eine Auswahl der zu diskutierenden Methoden getroffen. Weitgehend unberücksichtigt bleiben zwei wichtige Methodenklassen: (a) die verschiedenen Versionen des Rasch-Modells (s.a. Fischer, 1976, 1977, 1978) und (b) dynamische probabilistische Systeme (s. aber Levin & Burke, 1972; Bartholomew, 1973; Wiggins, 1973; Tack, 1976; Singer & Spilerman, 1976a,b; Lee, Judge & Zellner, 1977; Singer & Spilerman, 1979a,b; Tuma & Hannan, 1979; Wasserman, 1980; Tuma, 1980; Singer, 1981).

Die ausführliche Behandlung dynamischer probabilistischer Systeme würde einerseits den Umfang des Beitrages sprengen, andererseits das Problem aufwerfen, fast keine empirischen Referenzuntersuchungen nennen zu können. Das wird gerade an dem Artikel von Singer & Spilerman (1979a) in dem „neuen Klassiker“ der Veränderungsmessung von Nesselroade & Baltes (1979) deutlich.

Alle hier angeführten Beiträge: (a) Zeitreihenanalyse, (b) Zeitreihenexperimente, (c) Analyse von Differenzenwerten, (d) Strukturmodelle, (e) Varianzanalyse, (f) Markoffketten, (g) zeitkontinuierliche Modelle (Differentialgleichungssysteme, Markoffprozesse) werden nach mehreren Gesichtspunkten klassifiziert. Folgende Gesichtspunkte spielen dabei eine Rolle: (a) die Zeit wird diskret oder kontinuierlich angesehen, (b) uni- oder multivariate Methoden ( $M = 1$ ,  $M > 1$ ), (c)  $N = 1$  oder  $N > 1$  - Studien ( $N = \text{Zahl der Meßwertträger oder Pbn}$ ), (d) Test-Retest-Untersuchungen ( $T = 2$ ), Paneluntersuchungen ( $2 \leq T \leq 10$ ), Zeitreihenanalysen ( $50 \leq T$ ) (e) Experimente, Quasiexperimente, Feldstudien, (f) Meßniveau (dichotom, ordinal, Intervall).

Eine grobe Gliederung findet sich in Figur 1.1 (s. S. 244).

## 2. Univariate Zeitreihenanalyse

$N = 1$ ,  $M = 1$ ,  $T > 50$

Univariate Zeitreihenanalysen gehören noch nicht zum Standardinventar psychologischer Forschung, obwohl nur mit ihrer Hilfe bestimmte Fragestellungen befriedigend beantwortet werden können. So ist z.B. die Periodik von Stimmungszuständen von Wichtigkeit bei der Erforschung ihrer externen od. internen Auslösung. Eine ausführliche Diskussion der Bedeutung der Zeitreihenanalyse in Psychologie und Psychiatrie findet sich z.B. bei Chassan (1979<sup>2</sup>) und Gottman & Leiblum (1974).

Zeitreihenanalysemethoden unterscheiden sich hauptsächlich von anderen statistischen Methoden dadurch, daß sie die zeitliche Abhängigkeit von Beobachtungen berücksichtigen. Zeitlich frühere Beobachtungen dürfen im Modell

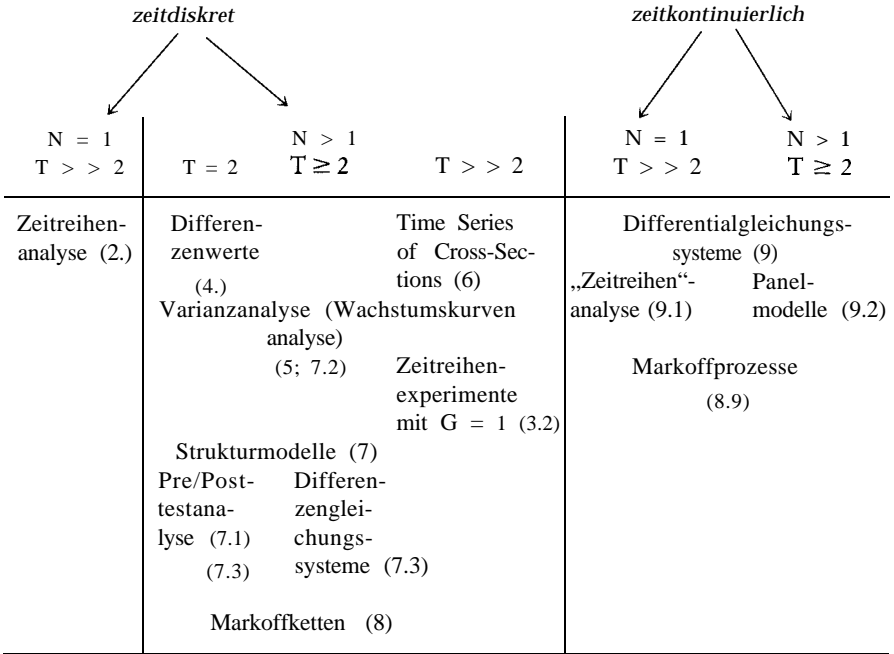


Fig. 1.1 : Gliederung der behandelten Themen im Bereich „Veränderungsmessung, -analyse und -prognose“

spätere Beobachtungen beeinflussen. Im Gegensatz zur Ökonomie konzentriert sich bisher in der Psychologie das Interesse an der Zeitreihenanalyse nicht auf Prognosen sondern mehr auf Interventionsevaluation und Analyse von Kovariationszusammenhängen. Bei der Interventionsevaluation wird untersucht, ob Eingriffe (Therapie, Arbeitsbeschaffung, Ortsveränderungen, Gesetze etc.) einen Einfluß auf abhängige Variable haben, die an dem Pbn über längere Zeit erhoben wurden (Gottman & Glass, 1978; Revenstorf, 1979; Revensdorf & Keeser, 1979; McCain & Cleary, 1979; Meier, 1981). Als Effekte können auftreten: Niveauänderungen der Zeitreihe, Trendumkehr bei Variablen der subjektiven Befindlichkeit u.a.m. Liegt das Interesse der Untersuchenden dagegen mehr auf Kovariationszusammenhängen, stellt sich die Frage, ob bestimmte unabhängige Variable als zeitlich führende Indikatoren für abhängige Variable angesehen werden können („lead and lag structure“). So können z.B. der Einfluß des Wetters oder der Jahreszeiten auf die Stimmung des Pbn (Zimmermann, 1979), die Abhängigkeit der Suizidneigung von der Arbeitslosigkeit (Vigderhous, 1978) oder so profane Fragen wie die lead/lag-Beziehung von Werbeintensität und Absatz (Helmer & Johansson, 1977) von Interesse sein.

Für beide Fragestellungen werden entweder *Interventionsmodelle* (dynamische Versionen der Varianzanalyse) oder *Transfermodelle* (dynamische Versionen der Regression) formuliert. Beide Modellklassen stützen sich dabei mathematisch auf die univariate Zeitreihenanalyse. Hierbei beschränken wir uns auf die ARIMA-Modelle von Box & Jenkins (1976<sup>3</sup>), die sich sowohl in theoretischer wie auch praktischer Hinsicht als sehr brauchbar (wenn auch etwas kompliziert in der Handhabung) herausgestellt haben.

Die ARIMA-Modelle setzen sich aus drei Prozeßkomponenten zusammen: (a) dem autoregressiven, (b) dem integrierenden und (c) dem moving-average Prozeß. Zum besseren Verständnis ihres Zusammenwirkens dient es, wenn man sich das Entstehen einer Zeitreihe nach dem ARIMA-Modell vergegenwärtigt. Man nimmt an, daß es eine Unzahl äußerer Einflüsse gibt, deren Resultante der „random shock“  $a_t$  ist. Diese äußeren Einflüsse wirken nun in bestimmter Art und Weise (auf eine Person) nach. Die Art des Nachwirkens wird modellgemäß durch die drei informationsverarbeitenden Prozesse (autoregressiver, integrierender und moving-average) abgebildet. Das Ergebnis dieses Signalverarbeitungsprozesses ist dann die Zeitreihe  $Y_t - L$  (mit einer Konstanten  $L$ , von der wir bis auf weiteres annehmen, daß sie gleich Null ist; diese Darstellungsform wird von Glass, Willson & Gottman (1975) gewählt; andere Darstellungsformen finden sich bei Möbus, Görlicke & Kröh, 1982) (Fig. 2.1a). Bei empirischen Untersuchungen kennt man dagegen die Entstehungsgeschichte der Zeitreihe und damit die Parameter der drei Prozesse nicht. Man versucht daher die Entstehungsgeschichte der Zeitreihe gewissermaßen „auf den Kopf zu stellen“ bzw. rückgängig zu machen und dadurch die Zeitreihe zu analysieren (Fig. 2.1 b).

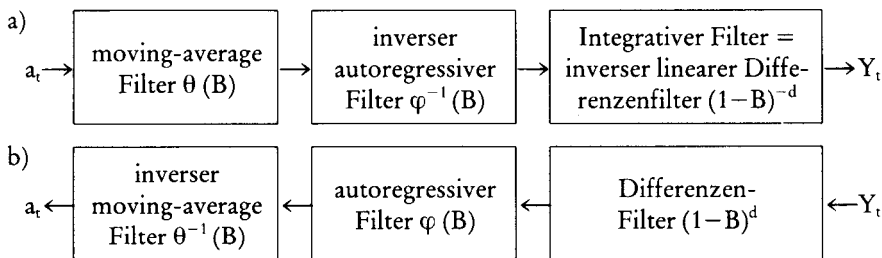


Fig. 2.1: (a) *Datengenerierung* mit dem ARIMA-Modell, (b) *Zeitreihenanalyse* als Prozeß der Wegfilterung systematischer Anteile aus der Zeitreihe  $Y_t$ , so daß nur der „white noise“-Prozeß  $a_t$  verbleibt.

Als Ergebnis einer adäquaten Datenanalyse verbleibt von der ursprünglich beobachteten Zeitreihe  $Y_t$  nur noch der generierende Zufallseinfluß  $a_t$ , der auch „weißes Rauschen“ genannt wird. Wir wollen im folgenden die drei Filter bzw. Prozesse beschreiben.

## 2.1 Integrierte Prozesse der Ordnung d: ARIMA(0,d,0)-Modelle

Ein Beispiel für einen integrierenden Prozeß ist der sogenannte „random walk“ :ARIMA(0,1,0). Der Name rührt von der Vorstellung, man würde den Nachhauseweg eines Betrunkenen beobachten. Viele kleine Anstöße  $a_t$  von rechts oder links erzeugen einen Weg, den man gemeinhin mit „Zick-Zack-Kurs“ bezeichnet. Der Abstand  $Y_t$  auf dem Bürgersteig zur rechten Begrenzung zum Zeitpunkt  $t$  setzt sich zusammen aus:

$$\begin{aligned}
 Y_0 &= a_0 & &= Y_0 & \text{dabei sollen sich die} \\
 Y_1 &= a_0 + a_1 & &= Y_0 + a_1 & \text{„random shocks“ unab-} \\
 Y_2 &= a_0 + a_1 + a_2 & &= Y_1 + a_2 & \text{hängig mit konstanter} \\
 & & & & \text{Varianz und Erwar-} \\
 & & & & \text{tungswert 0 normal} \\
 & & & & \text{verteilen} \\
 Y_t &= a_0 + a_1 + a_2 + \dots + a_t = Y_{t-1} + a_t & & a_t \sim \text{NID}(0, \sigma_a^2)
 \end{aligned}
 \tag{2.1}$$

Die Input-Output-Beziehung des „random walk“ läßt sich graphisch darstellen :

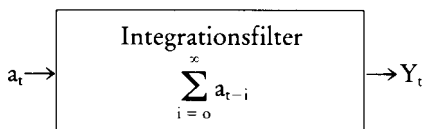


Fig. 2.2: Input-Output-Beziehung des „random-walk“

Im Random Walk ergibt sich  $Y_t$  als die Summe aller früheren Einflüsse. Der Prozeß integriert alle  $a_{t-i}$  zum Endresultat  $Y_t$ .

Leicht kommt es zu Fehlinterpretationen, wenn man den „Zick-Zack-Kurs“ des Prozesses nicht als stochastische Drift, die keine relevante Information bietet, sondern als deterministischen Trend oder gar als Effekt einer psychologischen Intervention (z.B. in einem ABAB Design) überinterpretiert. Zur Prüfung, ob ein Trend oder nur stochastische Drift vorliegt, muß man die Zeitreihe „differenzieren“. Dieses kann auch für die Schätzung der autoregressiven und moving-average-Filter notwendig werden. Die Differenzenbildung beseitigt den Trend im Niveau der Zeitreihe, wenn dieser nicht gerade exponentiell anwächst.

Im Random Walk mit Trend  $\theta_0$  verteilen sich die  $a_t \sim \text{NID}(\theta_0, \sigma_a^2)$ . Die Prozeßdarstellung für den ARIMA (0,1,0) verändert sich von (2.1) zu

$$\begin{aligned}
 (2.2) \quad & \begin{array}{rcl}
 & a_0^{\wedge} & \\
 Y_0 = (\theta_0 + a_0) & a_1^{\wedge} & = (\theta_0 + a_0) \\
 Y_1 = (\theta_0 + a_0) + (\theta_0 + a_1) & & = Y_0 + (\theta_0 + a_1) \\
 Y_2 = (\theta_0 + a_0) + (\theta_0 + a_1) + (\theta_0 + a_2) & & = Y_1 + (\theta_0 + a_2) \\
 \vdots & & \vdots \\
 Y_t = (\theta_0 + a_0) + \dots + (\theta_0 + a_t) & & = Y_{t-1} + (\theta_0 + a_t)
 \end{array}
 \end{aligned}$$

Bildet man jetzt die ersten Differenzen  $\Delta Y_t = Y_t - Y_{t-1}$

$$\begin{aligned}
 (2.3) \quad & \begin{array}{rcl}
 \Delta Y_1 = Y_1 - Y_0 & = & (\theta_0 + a_1) \\
 \Delta Y_2 = Y_2 - Y_1 & = & (\theta_0 + a_2) \\
 \vdots & & \\
 \Delta Y_t = Y_t - Y_{t-1} & = & (\theta_0 + a_t)
 \end{array}
 \end{aligned}$$

ist der Erwartungswert der Differenzen  $E[\Delta Y_t] = \theta_0$ . Zur Prüfung des Vorliegens eines Trends (hier linearer Trend), testen wir die Nullhypothese

$$H_0: \theta_0 = 0$$

Der in (2.2) dargestellte Random Walk (=ARIMA(0,1,0)) mit deterministischem Trend  $\theta_0$

$$(2.4a) \quad Y_t = Y_{t-1} + \theta_0 + a_t$$

läßt sich mit Hilfe des Backshiftoperators B schreiben als

$$(2.4b) \quad \Delta Y_t = (1 - B) Y_t = \theta_0 + a_t \quad \text{mit } B Y_t = Y_{t-1}$$

oder als

$$(2.4c) \quad (1 - B) (Y_t - L) = a_t \quad \text{mit } L = (1 - B)^{-1} \theta_0 = \sum_{i=0}^{\infty} \theta_0.$$

D.h. die Zeitreihe besitzt kein endliches Niveau L. Sie ist nichtstationär. Die verschiedenen Schreibweisen (2.4a)-(2.4c) sind näher bei Möbus, Görcke & Kröh (1982) erklärt.

Während die Rohwerte  $Y_t$  für alle Zeitpunkte kein festes Level besitzen (Nichtstationarität im homogenen Sinne), liegt dieses für die Differenzen  $\Delta Y_t = W_t$  vor. Deren Erwartungswert ist

$$(2.5) \quad E(\Delta Y_t) = \theta_0$$

Damit ist die Zeitreihe der Differenzen  $\Delta Y_t$  stationär im homogenen Sinn.

Folgt der Random Walk dagegen einer stochastischen Drift oder einem deterministischen Trend vom Grade  $d=2$ , muß das ARIMA(0,2,0)-Modell aufgestellt werden:

$$(2.6) \quad \Delta^2 Y_t = (1-B)^2 Y_t = \theta_0 + a_t \quad \text{mit} \quad \begin{cases} \Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1} = \\ (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ = Y_t - 2Y_{t-1} + Y_{t-2} = \\ (1 - 2B + B^2) Y_t = (1-B)^2 Y_t \end{cases}$$

Allgemein gilt für das ARIMA(0,d,0)-Modell

$$(2.7) \quad (1-B)^d Y_t = \theta_0 + a_t$$

Muß die Nullhypothese  $H_0: \theta_0 = 0$  verworfen werden, liegt ein deterministischer Trend vom Grade  $d$  vor (bei  $d=2$  eine Parabel). Kann die Nullhypothese beibehalten werden, haben wir es nur mit stochastischer Drift vom Grade  $d$  zu tun. Sie darf nicht inhaltlich interpretiert werden, weil sie nur ein Zufallsprodukt ist! Im ARIMA(0,d,0) ist zwar  $Y_t$  nichtstationär im homogenen Sinn, jedoch liegt die Stationarität im Level für die  $d$ -fachen Differenzen  $W_t$  vor:

$$E(\Delta^d Y_t) = E(W_t) = \theta_0$$

Für die Abbildung der Zeitreihe als autoregressiver oder moving average Prozeß wird noch die Stationarität in Varianz und Autokorrelation gefordert. Das bedeutet zum einen die Zeitunabhängigkeit der Varianz. Die Autokorrelation darf dagegen von zeitlichen Verschiebungen (Lags) zweier Zeitreihen abhängen. Liegt die Stationarität in der Varianz nicht vor, kann die Logarithmierung der eventuell vorher differenzierten Zeitreihe Abhilfe schaffen. Ein Beispiel findet sich bei McCain & McCleary (1979, S. 282-293). Eine der Varianzstationarität ähnliche Voraussetzung findet sich z.B. auch bei der Regression und ist dort als Homoskedastizitätsforderung bekannt.

## 2.2 Autoregressive Prozesse der Ordnung $p$ : ARIMA( $p,0,0$ ) u. ARIMA( $p,d,0$ )-Modelle

Die AR-Prozesse wurden von Yule (1926) eingeführt und generalisiert von Walker (1931). Man nimmt an, daß die Abweichungsvariable  $y_t = Y_t - L$  von  $p$  früheren Variablen und einem Random-Schock  $a_t$  abhängt:

$$(2.8a) \quad y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots + \varphi_p y_{t-p} + a_t$$

oder in Operatorschreibweise

$$(2.8b) \quad (1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p) y_t = a_t$$

oder ganz kurz

$$(2.8c) \quad \varphi(B) y_t = a_t$$

bzw.

$$(2.8d) \quad Y_t - L = \frac{1}{\varphi(B)} a_t = \varphi^{-1}(B) a_t$$

Da der AR-Prozeß stationär ist, müssen nichtstationäre Prozesse erst noch durch Differenzenbildung stationär gemacht werden, so daß sie in den Differenzen  $\Delta^d Y_t = (1 - B)^d Y_t = W_t$  definiert sind. Der ARIMA(p,d,0)-Prozeß läßt sich dann in Rohwerten  $Y_t$  (s.a. Hibbs, 1977, S. 140; Möbus, Görnicke & Kröh, 1982) formulieren als

$$\varphi(B)(1-B)^d Y_t = \theta_0 + a_t$$

$$(2.9a) \quad Y_t = \frac{\theta_0 + a_t}{\varphi(B)(1-B)^d} \quad \text{mit } \theta_0 = L \varphi(B)(1-B)^d$$

Andere Autoren (z.B. Jenkins, 1979, S. 98) wählen eine andere Schreibweise für den gleichen Sachverhalt. Sie formulieren den Prozeß als ARIMA(p,0,0)-Modell in den Abweichungen der Differenzen vom zugehörigen Differenzmittelwert  $W_t - \mu_W$ :

$$\text{mit } W_t = (1-B)^d Y_t$$

$$(2.9b) \quad W_t - \mu_W = \frac{1}{\varphi(B)} a_t \quad \mu_W = E(W_t) = \frac{\theta_0}{\varphi(B) \cdot 1}$$

Am häufigsten tritt der ARIMA(1,0,0) auf:

$$(2.10a) \quad y_t = \varphi_1 y_{t-1} + a_t \text{ mit den Stationaritätsgrenzen } -1 < \varphi_1 < +1$$

oder

$$(2.10b) \quad (1 - \varphi_1 B) y_t = a_t$$

Durch sukzessives Einsetzen kann man  $y_t$  ähnlich wie bei (2.1) auf die Random Shocks der Vergangenheit zurückführen:

$$\begin{aligned} y_0 &= a_0 \\ y_1 &= \varphi_1 a_0 + a_1 \\ y_2 &= \varphi_1^2 a_0 + \varphi_1 a_1 + a_2 \\ &\dots \end{aligned}$$



$$\begin{aligned}
 (2.11) \quad y_t &= \varphi_1^t a_0 + \varphi_1^{t-1} a_1 + \dots + \varphi_1 a_{t-1} + a_t = \sum_{i=0}^{\infty} \varphi_1^i a_{t-i} \\
 &= (1 + \varphi_1 B + \varphi_1^2 B^2 + \dots + \varphi_1^k B^k + \dots) a_t \\
 &= (1 - \varphi_1 B)^{-1} a_t \\
 &= \frac{1}{\varphi(B)} a_t
 \end{aligned}$$

wenn man wie Glass et al. (1975, S. 90) die  $a_t$  mit Index  $t < 0$  auf ihren Erwartungswert  $E(a_t) = 0$  setzt

Dabei ist die Inverse des Differenzenoperators  $(1-B)$  definiert

$$\text{als } (1-B)^{-1} = 1 + B + B^2 + B^3 + \dots = \sum_{k=0}^{\infty} B^k. \quad \text{Damit ist}$$

$$(1 - \varphi_1 B)^{-1} = 1 + \varphi_1 B + \varphi_1^2 B^2 + \dots = \sum_{k=0}^{\infty} \varphi_1^k B^k.$$

Ein Vergleich mit (2.11) zeigt, daß man auf einfache Weise die Differenzengleichung (2.10) lösen kann zu:

$$Y_t - L = \frac{1}{\varphi(B)} a_t \quad \text{mit } L = \mu_y$$

was der Form (2.9b) entspricht. Da der  $|\varphi_1| < 1$  ist, findet im Gegensatz zum Random Walk eine exponentielle Abschwächung äußerer Einflüsse statt: für  $t \rightarrow \infty$  geht  $\varphi_1^t \rightarrow 0$ , d.h. der Anfangseinfluß  $a_0$  spielt im ARIMA (1,0,0) kaum noch eine Rolle. Ein Beispiel für einen ARIMA(1,0,0) findet sich in Fig. 2.5.

Im Gegensatz zum Random Walk (Fig. 2.2) hat das Input-Output-Modell des ARIMA (1,0,0) ein „Leck“, durch das frühere äußere Einflüsse  $a_{t-i}$  verschwinden. Das Input-Output-Modell des ARIMA (1,d,0)-Modells sieht dann so aus (Fig. 2.3):

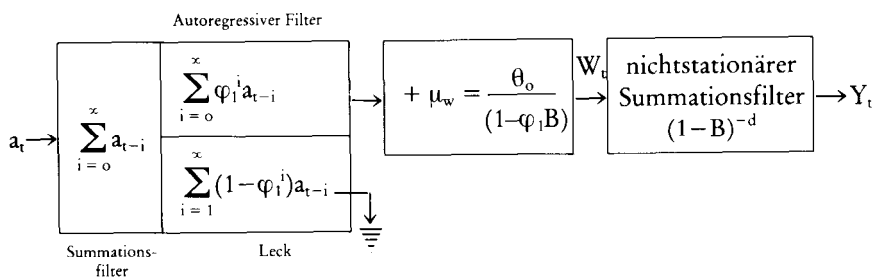


Fig. 2.3: Input-Output-Beziehung des ARIMA (1,d,0)-Prozesses

Ebenfalls relativ häufig kommt noch der ARIMA (2,0,0)-Prozeß vor:

$$(2.12a) \quad y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + a_t \quad \text{mit den Stationaritätsgrenzen} \\ \text{(s.a. Fig. 2.4):}$$

$$\text{oder} \quad -1 < \varphi_2 < +1$$

$$(2.12b) \quad (1 - \varphi_1 B - \varphi_2 B^2) y_t = a_t \quad \varphi_1 + \varphi_2 < +1$$

$$\text{bzw.} \quad \varphi_2 - \varphi_1 < +1$$

$$(2.12c) \quad Y_{t-L} = \frac{1}{(1 - \varphi_1 B - \varphi_2 B^2)} a_t = \frac{1}{\varphi(B)} a_t$$

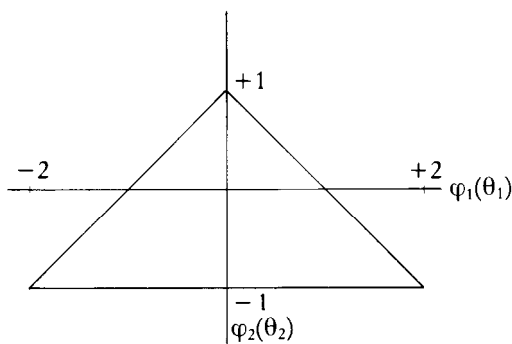


Fig. 2.4: Stationaritätsgrenzen für den ARIMA(2,0,0) bzw. Invertierbarkeitsgrenzen (s.U.) für den ARIMA(0,0,2)

## 2.3 Moving-average Prozesse der Ordnung q: ARIMA(0,0,q)-Modelle

Die moving-average-Prozesse sind auch schon relativ lange bekannt. Sie gehen auf Slutsky (1937) zurück. Während bei den integrierenden Prozessen die früheren random shocks  $a_t$  ohne Abschwächung ihren Einfluß theoretisch unendlich lange ausüben, wird ihr Einfluß im autoregressiven Prozeß mit der Zeit allmählich schwächer. Beim moving-average Prozeß dagegen hält sich der Einfluß nur  $q$  Zeitpunkte, um dann plötzlich ganz zu verschwinden. Die MA-Prozesse sind immer stationär.

Der einfachste Prozeß ist der ARIMA(0,0,1):

$$(2.13a) \quad y_t = a_t - \theta_1 a_{t-1} \quad \text{mit der Invertierbarkeitsbedingung } -1 < \theta_1 < +1$$

oder

$$(2.13b) \quad y_t = (1 - \theta_1 B)a_t$$

oder ganz kurz

$$(2.13c) \quad y_t = \theta(B)a_t \text{ oder } Y_t - L = \theta(B)a_t \quad \text{mit } L = \theta_0$$

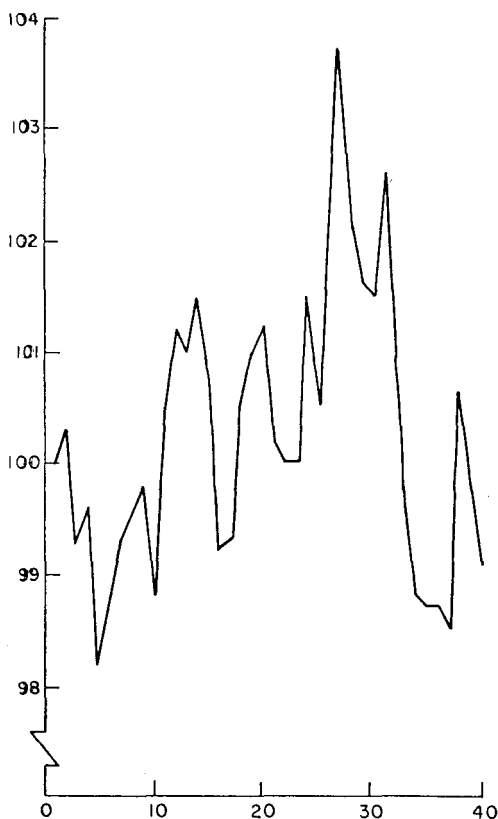


Fig. 2.5: Beispiel für einen ARIMA (1,0,0)-Prozeß (aus Glass et al., 1975)

Unter Invertierbarkeit versteht man die Forderung, daß die Parameter mit der die Vergangenheit einer Zeitreihe gewichtet wird, um Prognosen zu erhalten, mit zunehmendem Zurückgehen in die Vergangenheit allmählich verschwinden (s. Jenkins, 1979, S. 134).

Das Input-Output-Modell des ARIMA(0,0,1) findet sich in Fig. 2.6.

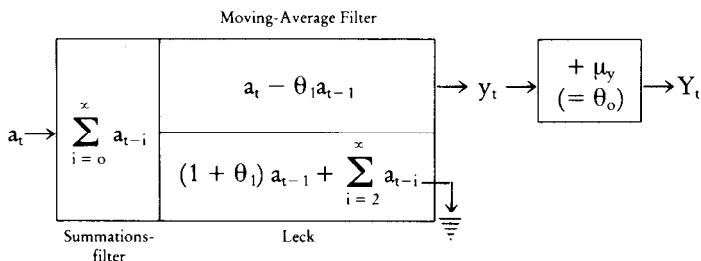


Fig. 2.6: Input-Output-Modell des ARIMA (0,0,1)

Durch sukzessive Substitution von (2.13a) kann man zeigen, daß der ARIMA(0,0,1) ein unendlich langer autoregressiver Prozeß mit einer ganz einfachen Parameterstruktur ist:

$$(2.14) \quad y_t = a_t - \sum_{i=1}^{\infty} \theta_1^i y_{t-i}$$

Ein Beispiel für einen ARIMA(0,0,1) ist in Fig. 2.7 dargestellt.

Der ARIMA(0,0,2)-Prozeß läßt sich analog zu (2.13) schreiben:

$$(2.15a) \quad y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

$$(2.15b) \quad y_t = (1 - \theta_1 B - \theta_2 B^2) a_t$$

$$(2.15c) \quad Y_t - L = \theta(B) a_t$$

mit Invertierbarkeitsgrenzen für  $\theta_1$  und  $\theta_2$ , die numerisch den Stationaritätsgrenzen des ARIMA(2,0,0) entsprechen (s. Fig. 2.4)

Nur wenn die Invertierbarkeitsbedingung erfüllt ist, läßt sich der MA-Prozeß als unendlicher AR-Prozeß formulieren. Sollten *Parameterschätzungen* außerhalb der Stationaritäts- bzw. der Invertierbarkeitsgrenzen liegen, ist das ein sicheres Zeichen, daß die Zeitreihe nicht stationär ist und differenziert werden muß oder daß zu häufig differenziert wurde.

## 2.4 Das allgemeine ARIMA (p,d,q)-Modell

Der Hauptgrund für die Kombination der drei Prozesse in ein Modell liegt im Bestreben mit möglichst wenigen Parametern möglichst komplizierte Zeitreihen beschreiben zu können. Besonders wenn man ein gemischtes Modell (sowohl p als auch q ungleich Null) formuliert, sollte man sich jedoch gegen die

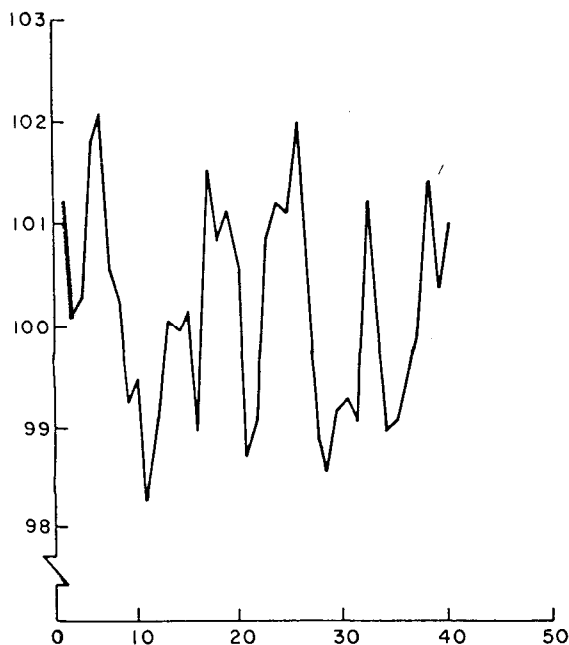


Fig. 2.7: Beispiel für einen ARIMA (0,0,1)-Prozeß (aus Glass et al., 1975)

Gefahr der Überparametrisierung schützen (Box & Jenkins, 1976, Kap. 7.3.5). Dieses soll am einfachsten gemischten Prozeß dem ARIMA(1,0,1) gezeigt werden.

$$(2.16a) \quad y_t = \varphi_1 y_{t-1} + a_t - \theta_1 a_{t-1}$$

$$(2.16b) \quad (1 - \varphi_1 B)y_t = (1 - \theta_1 B)a_t$$

$$(2.16c) \quad y_t = (1 - \varphi_1 B)^{-1}(1 - \theta_1 B) a_t$$

$$(2.16d) \quad Y_t - L = \frac{\theta(B)}{\varphi(B)} a_t$$

mit den kombinierten Stationaritäts- und Invertierbarkeitsbedingungen -  $1 < \varphi_1 < 1$  und -  $1 < \theta_1 < 1$  (s.a. Fig. 2.8) (S. 255).

Ist nun  $\varphi_1 = \theta_1$  läßt sich der Prozeß exakt auf den ARIMA(0,0,0), nämlich das weiße Rauschen, reduzieren:

$$(2.17) \quad y_t = a_t$$

Bei anderen Parameterkonstellationen sind annähernde Reduktionen auf einen ARIMA (0,0,1), einen ARIMA(0,0,2) oder einen ARIMA(1,0,0) möglich (Box & Jenkins, 1976, 248-250).

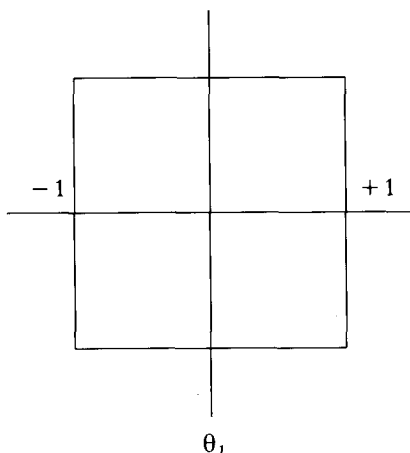


Fig. 2.8: Stationaritäts- und Invertierbarkeitsbedingungen für den ARIMA (1,0,1)

Ist man dagegen der Auffassung, daß die Zeitreihe einem vollständigen ARIMA(p,d,q)-Modell entspricht und sich dieses nicht vereinfachen läßt, können wir entsprechend (2.9a) und (2.9b) zwischen zwei äquivalenten Darstellungsformen wählen.

Beschreiben wir den Prozeß in Rohwerten  $Y_t$  als ARIMA(p,d,q), haben wir:

$$(2.18a) \quad (1 - \varphi_1 B - \dots - \varphi_p B^p) (1 - B)^d Y_t = \theta_0 + (1 - \theta_1 B - \dots - \theta_q B^q) a_t$$

oder

$$(2.18b) \quad \varphi(B) (1 - B)^d Y_t = \theta_0 + \theta(B) a_t$$

bzw.

$$(2.18c) \quad Y_t = \frac{\theta_0 + \theta(B)}{\varphi(B)(1 - B)^d} a_t$$

Diese Darstellungsform präferiert Hibbs (1977, S. 140). Die Alternative hierzu ist die Erweiterung zu (2.9b) mit

$$(2.19) \quad W_t - \mu_W = \frac{\theta(B)}{\varphi(B)} a_t \quad \begin{array}{l} \text{mit } W_t = (1 - B)^d Y_t \\ \text{und } \mu_W = \frac{\theta_0}{\varphi(B) \cdot 1} \end{array}$$

Diese Darstellung findet sich bei Jenkins (1977, S. 98). Sind die Parameter  $\theta_0$  oder  $\mu_W$  von Null verschieden, liegt wieder ein deterministischer Trend vom Grade d vor.

Ein Beispiel für einen  $ARIMA(0,2,2)$  sieht man in Fig. 2.9. Zu beachten ist der quadratische Verlauf, dem man nicht ansehen kann, ob er stochastische Drift oder einen deterministischen Trend enthält. Das kann wie oben gezeigt nur im Wege der wiederholten Differenzenbildung und der anschließenden statistischen Prüfung des Parameters  $\theta_0$  bzw.  $\mu_w$  genauer untersucht werden.

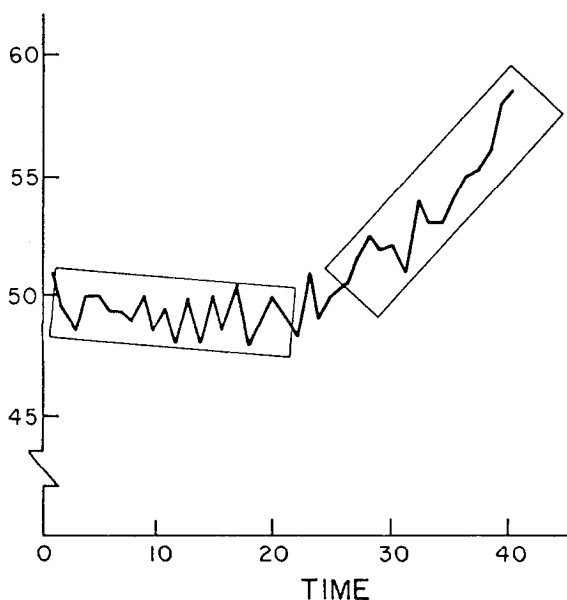

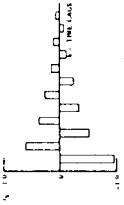
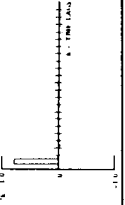
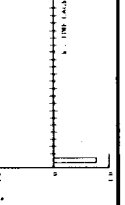
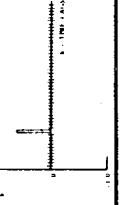


Fig. 2.9: Beispiel für einen nichtstationären  $ARIMA(0,2,2)$

## 2.5 Autokorrelations- und partielle Autokorrelationsfunktion

Einer empirisch gewonnenen Zeitreihe sieht man die Grade  $p, d, q$  des  $ARIMA$ -Modells nicht an. Daher muß in einem iterativen Identifikationsprozeß die Festlegung der  $p, d, q$  vor der Parameterschätzung erfolgen. Die Identifikation stützt sich dabei auf die Autokorrelationsfunktion (ACF) und die partielle Autokorrelationsfunktion (PACF). Jeder stationäre Prozeß besitzt eine spezielle unverwechselbare Form beider Funktionen. Werden sie dagegen mit Daten geschätzt, versucht man die empirisch erhaltenen Muster mit den theoretischen zu vergleichen und damit die Grade  $p, d, q$  des  $ARIMA$ -Modells festzulegen.

Die Autokorrelationsfunktion ist definiert als

Autocorrelation Function (ACF)	Description of ACF	Tentative Model	Estimated Coefficient
	Decays exponentially	Autoregressive First Order $(1 - \phi_1 B) \tilde{z}_t = a_t$ $\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + a_t$	$\phi_1 > 0$
	Decays exponentially in oscillation		$\phi_1 < 0$
	Spikes at lag 1	Moving-Average First Order $\tilde{z}_t = (1 - \theta_1 B) a_t$ $= a_t - \theta_1 a_{t-1}$	$\theta_1 < 0$
			$\theta_1 > 0$
	Spikes at lag 12	Seasonal Moving Average $\tilde{z}_t = (1 - \theta_{12} B^{12}) a_t$ $= a_t - \theta_{12} a_{t-12}$	$\theta_{12} < 0$



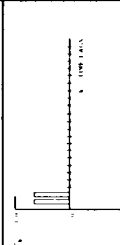
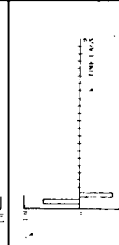
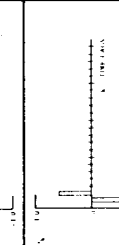
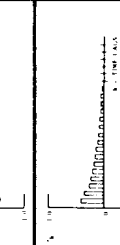
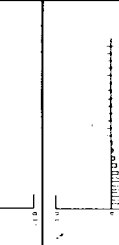
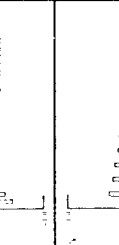

	Spikes at lags 1 and 2	Moving-Average Second Order $\hat{z}_t = (1 - \theta_1 B - \theta_2 B^2) a_t$ $= a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$	$\theta_1 > 0 \quad \theta_2 > 0$	
			$\theta_1 > 0 \quad \theta_2 < 0$	
	Spikes at lags 1 and 2		$\theta_1 < 0 \quad \theta_2 > 0$	
			$\theta_1 < 0 \quad \theta_2 < 0$	
	Decays exponentially from lag 1	Mixed ARMA—(1, 0, 1) $(1 - \phi_1 B) \hat{z}_t = (1 - \theta_1 B) a_t$ $\hat{z}_t - \phi_1 \hat{z}_{t-1} = a_t - \theta_1 a_{t-1}$ $\hat{z}_t = \phi_1 \hat{z}_{t-1} + a_t - \theta_1 a_{t-1}$	$(\phi_1 - \theta_1) > 0 \quad \phi_1 > 0$	
			$(\phi_1 - \theta_1) < 0 \quad \phi_1 > 0$	
	Decays exponentially from lag 1 in oscillation		$(\phi_1 - \theta_1) < 0 \quad \phi_1 < 0$	
				

Fig. 2.10: Theoretische Autokorrelationsfunktionen (ACFs) für verschiedene stationäre Prozesse (dem  $Z$  und  $\Phi$  in der Figur entsprechen  $y$  und  $\varphi$  im Text)

$$(2.20) \quad \text{ACF}(k) = \varrho(k) = \frac{\gamma(k)}{\gamma(0)} \quad \begin{array}{l} k = \text{Verschiebung der Zeit-} \\ \text{reihe um } k \text{ Zeitpunkte} \\ (= \text{Lag}) \end{array}$$

mit der Kovarianz

$$\gamma(k) = E[(Y_t - \mu_y)(Y_{t+k} - \mu_y)]$$

Die Autokorrelation mißt also die Größe des Zusammenhangs der Zeitreihe mit der um  $k$  Zeitpunkte verschobenen. Weist z.B. eine Person einen ausgeprägten wöchentlichen Rythmus im Stimmungsbild auf (montags: blau; dienstags, mittwochs: mittelpträchtig; Donnerstag, Freitag: Vorfreude aufs Wochenende; Samstag und Sonntag: Hochstimmung), wird die Autokorrelationsfunktion für den Lag  $k=7$  einen ausgeprägt hohen Wert annehmen, da sich das Stimmungsbild von Montag zu Montag, von Dienstag zu Dienstag . . . ähnelt bzw. wiederholt (s.a. Huba, Lawley, Stallone & Fieve, 1976).

Ein gebräuchlicher Schätzer für die Autokorrelationsfunktion ist:

$$r(k) = \frac{c(k)}{c(0)}$$

mit Kovarianz  $c(k) = \frac{1}{N^*} \sum_{t=1}^{N-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})$

und  $N^* = N$  (Box & Jenkins, 1976)  
 $N^* = N-k$  (McCleary & Hay, 1980)

Mit zunehmend größerem Lag  $k$  wird dabei die Schätzung unsicherer, da die Zahl der Kreuzprodukte in der Kovarianzformel  $c(k)$  mit Wachsändern  $k$  ständig abnimmt.

Wir wollen kurz die ACFs einiger Prozesse beschreiben. Für den ARIMA (0,0,0) („weißes Rauschen“)

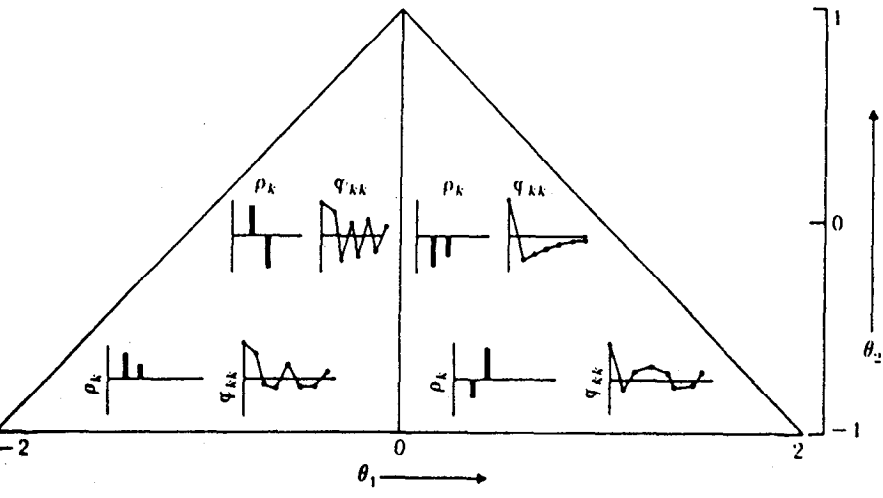
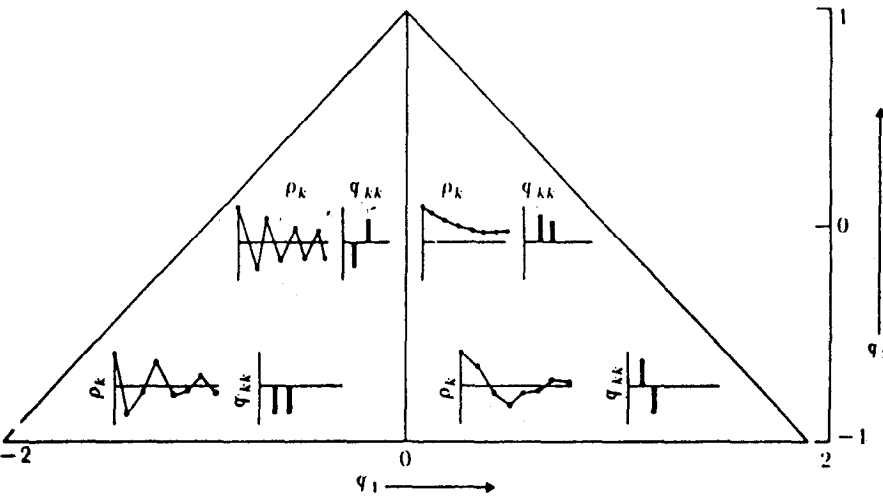
$$Y_t = a_t + \theta_0$$

ist die ACF  $(k) = 0$  für alle  $k$  größer als 0. Für den ARIMA (0,1,0)

$$(1 - B) Y_t = a_t + \theta_0$$

fällt die ACF ganz langsam ab. Denn wenn z.B.  $Y_t < Y_{t+k}$  für jedes  $k$ , wird die Kovarianz immer recht hoch sein. Ein ähnlich langsamer Abfall ist für alle nichtstationären ARIMA  $(p,d,q)$  zu erwarten.

Etwas schneller fällt die ACF( $k$ ) bei den autoregressiven Prozessen ab. So weist der ARIMA(1,0,0)  $(1 - \varphi_1 B)y_t = a_t$  eine sich exponentiell reduzierende ACF auf, wenn  $\varphi_1$  größer Null ist (Fig. 2.10)! Die Schnelligkeit der Reduktion hängt von der Größe von  $\varphi_1$  ab. Je größer  $\varphi_1$  ist, desto langsamer fällt die ACF, da ja nach (2.10)  $y_t$  dem  $y_{t-1}$  immer ähnlicher wird, je mehr  $\varphi_1$  an 1



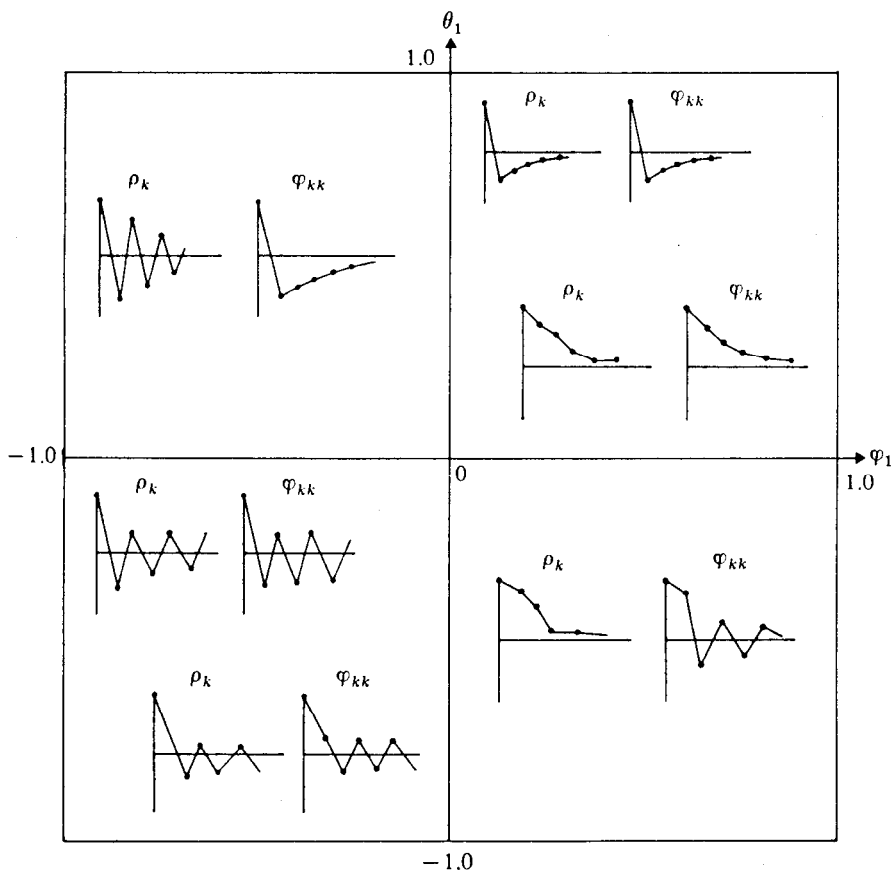


Fig. 2.1.1: ACF und PACF für verschiedene einfache stationäre Prozesse

herangeht. Ist  $\varphi_1 = 1$ , bleibt die ACF konstant. Wir haben es dann mit einem ARIMA (0,1,0) zu tun. Höhere ARIMA(p,0,0) Prozesse haben ebenfalls eine abfallende ACF, deren Muster jedoch auf recht komplizierte Weise von den  $\varphi$ -Parametern abhängen.

Bei den ARIMA(0,0,q) ist dagegen die ACF(k) für  $q < k$  ganz plötzlich Null, was die Identifikation der moving average-Prozesse durch die empirisch geschätzte ACF ganz wesentlich erleichtert. Eine Auswahl von theoretischen ACFs für verschiedene stationäre ARIMA-Prozesse findet sich in Fig. 2.10.

Leider erkennt man beim Studium der Fig. 2.10, daß nur der Grad  $q$  der reinen Moving-average-Prozesse leicht identifiziert werden kann. Gemischte ARIMA(p,0,q) und ARIMA(p,0,0) lassen sich mit Hilfe der an Daten geschätzten

ACF nicht auseinanderhalten. Ebenso ist es fast unmöglich, den Grad  $p$  bei reinen autoregressiven  $\text{ARIMA}(p,0,0)$ -Prozessen allein mit der geschätzten ACF festzulegen.

Glücklicherweise hilft in diesen nicht gut entscheidbaren Fällen die partielle Autokorrelationsfunktion  $\text{PACF}(k)$  weiter. Die Prozeßgleichungen für verschieden lange  $\text{ARIMA}(k,0,0)$  Modelle enthalten als letzten Koeffizienten  $\varphi_{kk}$  die partielle Autokorrelation, so:

(2.21a)  $\varphi_{11}$  für  $\text{ARIMA}(1,0,0)$   $y_t = \varphi_{11}y_{t-1} + a_t$

(2.21b)  $\varphi_{22}$  für  $\text{ARIMA}(2,0,0)$   $y_t = \varphi_{21}y_{t-1} + \varphi_{22}y_{t-2} + a_t$

(2.21c)  $\varphi_{33}$  für  $\text{ARIMA}(3,0,0)$   $y_t = \varphi_{31}y_{t-1} + \varphi_{32}y_{t-2} + \varphi_{33}y_{t-3} + a_t$   
.....  
usw.

Die partielle Autokorrelation  $\varphi_{kk}$  spiegelt den direkten Einfluß des Zeitpunkts  $t-k$  auf  $t$  wieder. Sammelt man die  $\varphi_{11}, \varphi_{22}, \varphi_{33} \dots$ , erhält man die partielle Autokorrelationsfunktion. Für einen autoregressiven Prozeß der Ordnung  $p$  werden die  $|\varphi_{kk}| > 0$  für  $k \leq p$  und  $|\varphi_{kk}| = 0$  für  $p < k$  sein. Die PACF hat also für die Identifikation die angenehme Eigenschaft, nach Lag  $p$  plötzlich Null zu werden.

Empirisch wird die  $\text{PACF}(k)$  durch sukzessives Anpassen von  $\text{ARIMA}(1,0,0)$ ,  $\text{ARIMA}(2,0,0)$ ,  $\text{ARIMA}(3,0,0) \dots$  Modellen und das Lösen der Yule-Walker-Gleichungen (2.22a) geschätzt (Box & Jenkins, 1976, 64-84). Diese Gleichungen besitzen eine Parallele bei der Bestimmung von Regressionsgewichten in der multiplen Regression (2.22b) (s.a. Cooley & Lohnes, 1971, S. 53 und Tab. 2.1)

Tabelle 2.1a: YULE-WALKER-Gleichungen zur Bestimmung der Parameter eines  $\text{ARIMA}(k, 0, 0)$  und der partiellen Autokorrelation  $\varphi_{kk}$

Tabelle 2.1b: Bestimmungsgleichung für die Regressionsgewichte  $\beta$  der multiplen Regression von  $k$  Prediktoren  $X_1, \dots, X_k$  auf das Kriterium  $Y$

$\begin{bmatrix} \hat{\varphi}_{k1} \\ \hat{\varphi}_{k2} \\ \vdots \\ \hat{\varphi}_{kk} \end{bmatrix} = \begin{bmatrix} 1 & r_1 & r_2 & \dots & r_{k-1} \\ r_1 & 1 & r_1 & \dots & r_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{k-1} & r_{k-2} & r_{k-3} & \dots & 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix}$   
(2.22a)  $\hat{\varphi}_k = \hat{P}_k^{-1} r_k$

$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \dots & 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} r_{1Y} \\ r_{2Y} \\ \vdots \\ r_{kY} \end{bmatrix}$   
(2.22b)  $\hat{\beta} = \hat{R}_{XX}^{-1} r_{XY}$

Die Schätzung der  $\text{PACF}(k)$  wird durch Einsetzen der geschätzten Autokorrelationen  $\hat{\varphi}_k = r_k$  in (2.22a) und systematischer Durchrechnung der Gleichungen  $\hat{\varphi}_k = \hat{P}_k^{-1} r_k$  ( $k=1,2,\dots,p,p+1,\dots$ ) gewonnen. Der Vergleich der geschätz-

ten PACF mit den theoretischen (s. Fig. 2.11) ermöglicht dann die Bestimmung der Ordnung  $p$  des  $\text{ARIMA}(p,0,0)$ -Modells, weil man nach lag  $p$  einen scharfen Knick in der PACF erwartet. Alle  $\text{PACF}(k)$  mit einem größeren Lag als  $p$  sollen dann annähernd Null sein (Beispiele finden sich in Fig. 2.11).

## 2.6 Saisonale Einflüsse

Wir werden in der Psychologie relativ selten „saisonale“ Einflüsse bei den Zeitreihen berücksichtigen müssen. Liegen dagegen bei täglicher Messung ausgeprägte Wochen- oder Monatszyklen vor, werden die Autokorrelations- oder partiellen Autokorrelationsfunktionen beim Lag 7 oder Lag 30 zusätzlich hohe Werte aufweisen (s.a. Fig. 2.10 für die ACF eines saisonalen Moving Average Prozesses). Wird z. B. ein  $\text{ARIMA}(1,0,0)$ -Prozeß von einem autoregressiven wöchentlichen Rhythmus vom Grade 2 überlagert, wird  $y_{21}$  auf vielfältige Weise beeinflusst (Fig. 2.12).

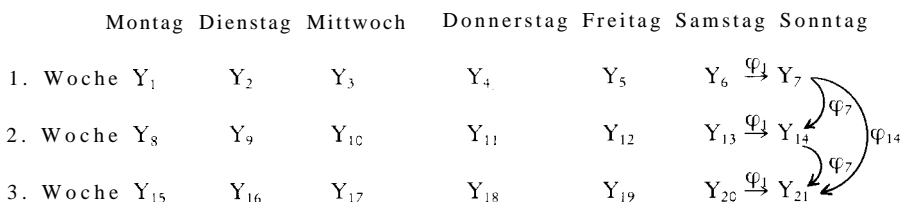


Fig. 2.12: Beeinflussungsstruktur bei einem  $\text{ARIMA}(1,0,0)$   $(2,0,0)_7$ -Prozeß

Neben der direkten Beeinflussung von  $y_{21}$  durch  $y_{20}$  liegt eine wöchentliche Ausstrahlung von  $y_{14}$  auf  $y_{21}$  und von  $y_7$  auf  $y_{14}$  sowie auf  $y_{21}$  vor. Wir können das normale  $\text{ARIMA}(1,0,0)$ -Modell

$$(2.23) \quad (1 - \varphi_1 B)y_t = \varepsilon_t \quad (\varepsilon_t \text{ ist kein weißes Rauschen})$$

um den Wochenzyklus

$$(2.24) \quad (1 - \varphi_7 B^7 - \varphi_{14} B^{14})\varepsilon_t = a_t$$

zwar nicht additiv, sondern multiplikativ erweitern zum  $\text{ARIMA}(1,0,0)$   $(2,0,0)_7$

$$(2.25a) \quad (1 - \varphi_1 B)(1 - \varphi_7 B^7 - \varphi_{14} B^{14})y_t = a_t$$

oder ausmultipliziert

$$(2.25b) \quad (1 - \varphi_7 B^7 - \varphi_{14} B^{14} - \varphi_1 B + \varphi_1 \varphi_7 B^8 + \varphi_1 \varphi_{14} B^{15}) y_t = a_t$$

gibt

$$(2.25c) \quad y_t = \varphi_1 y_{t-1} + \varphi_7 y_{t-7} + \varphi_{14} y_{t-14} - \varphi_1 \varphi_7 y_{t-8} - \varphi_1 \varphi_{14} y_{t-15} + a_t$$

und für  $y_{21}$

$$(2.26) \quad y_{21} = \varphi_1 y_{20} + \varphi_7 y_{14} + \varphi_{14} y_7 - \varphi_1 \varphi_7 y_{13} - \varphi_1 \varphi_{14} y_6 + a_{21}$$

Ein Vergleich mit Fig. 2.12 zeigt die Berechtigung dieses multiplikativen Ansatzes, weil hier ähnlich wie in der Pfadanalyse der Effekt eines Pfades der Länge  $k$  durch das Produkt der Pfadkoeffizienten  $\prod_{i=1}^k \varphi_i$  ausgedrückt wird.

In ähnlicher Weise lassen sich „saisonale“ integrierte und moving-average Prozesse zum allgemeinen  $\text{ARIMA}(p,d,q)_s$  ( $P,D,Q$ )<sub>s</sub> (Box & Jenkins, 1976, Kap. 9) anfügen, wobei  $s$  die Spanne des „saisonalen“ Effekts angibt (hier war  $s=7$  Tage). Wie man an (2.25) sieht, ist die Identifikation des Prozesses wegen der Produktterme  $\varphi_1 \varphi_7 y_{13}$  und  $\varphi_1 \varphi_{14} y_6$  schwieriger geworden als es bei einem reinen  $\text{ARIMA}(p,d,q)$  oder  $\text{ARIMA}(P,D,Q)$  der Fall gewesen wäre.

## 2.7 Modellidentifikation

Bei der Modellidentifikation sollte man ein möglichst einfaches Modell im Auge halten, um der Gefahr der Überparametrisierung zu entgehen. So wird aus einem  $\text{ARIMA}(0,0,1)$ -Modell  $y_t = (1 - \theta_1 B) a_t$  bei zu häufiger Differenzbildung der Zeitreihe ein überkompliziertes  $\text{ARIMA}(0,0,2)$ -Modell:

$$(2.27a) \quad \Delta y_t = a_t - (1 + \theta_1) a_{t-1} + \theta_1 a_{t-2}$$

oder

$$(2.27b) \quad \Delta y_t = (1 - (1 + \theta_1) B + \theta_1 B^2) a_t$$

Der Prozeß der Identifikation beginnt mit einem Plot der Zeitreihe, um eine eventuell vorliegende Nichtstationarität der Varianz zu entdecken. Diese wird z.B. durch Logarithmierung und Differenzenbildung beseitigt (s.a. Jenkins, 1979), wenn die Varianz proportional zum Level zunimmt (McCleary & Hay, 1980, S. 52). Allerdings verändert sich der Charakter des Modells. So wird aus einem in der logarithmierten Form additiven Random Walk in der Originalmetrik  $Y_t$  ein multiplikativer. Liegt kein Verdacht auf Nichtstationarität der Varianz vor, wird die ACF inspiziert. Fällt sie auch bei großen Lags  $k$  kaum ab, liegt der Verdacht auf Nichtstationarität im Level und einen  $\text{ARIMA}(p,d,q)$ -Prozeß nahe. Es muß dann die Zeitreihe  $Y_t$   $d$ -fach differenziert

werden. Anschließend werden von der (eventuell differenzierten) Zeitreihe die Muster der ACF mit den theoretisch zu erwartenden verglichen. Zur Beurteilung, wo Knicke in der ACF und der PACF auftreten, kann man die Standardschätzfehler und approximativen Konfidenzintervalle heranziehen.

Nach Quenouille (1949) ist unter der Hypothese, daß ein ARIMA(p,0,0) vorliegt, der Standardschätzfehler der partiellen Autokorrelation höherer Lags

$$(2.28) \quad \hat{\sigma} [\hat{\phi}_{kk}] \approx \left[ \frac{1}{T} \right]^{1/2} \quad p < k$$

d.h. ein approximatives Konfidenzintervall unter dieser Hypothese läßt sich mit  $\pm 2\hat{\sigma}[\hat{\phi}_{kk}]$  angeben. Als Regel für die Datenanalyse läßt sich formulieren: fällt ein  $\hat{\phi}_{kk}$  aus dem Konfidenzbereich, ist zu vermuten, daß  $p = k$  ist.

Der approximative Standardschätzfehler für die Autokorrelation  $\hat{Q}_k$  ist für einen ARIMA(0,0,q):

$$(2.29) \quad \hat{\sigma}[\hat{Q}_k] \approx \left[ \frac{1}{T} \left( 1 + 2 \sum_{v=1}^q \hat{Q}_v^2 \right) \right]^{1/2} \quad q < k$$

Ein approximatives Konfidenzintervall unter dieser Hypothese läßt sich ebenfalls mit  $\pm 2\hat{\sigma}[\hat{Q}_k]$  angeben. Fällt ein  $\hat{Q}_k$  aus dem Konfidenzbereich, ist zu vermuten, daß  $q = k$  ist, d.h.  $q$  war zu klein gewählt. Dabei müssen aber immer Unsicherheiten in Kauf genommen werden, die sich durch die Diskrepanz von geschätzten und theoretischen ACFs und PACFs ergeben. Diese Diskrepanz kann, wie Monte-Carlo-Studien zeigen, (Nelson, 1973, S. 75) recht groß werden. Daher gehen in die Interpretation stark subjektive Momente ein (Chatfield, 1975, S. 25).

Der vorläufigen Identifikation schließt sich die Schätzung der Parameter an. Zwar geben Box & Jenkins (1976, 517-520) Nomogramme an, nach denen man bei Kenntnis der Autokorrelationen die Parameter einiger einfacher Prozesse ablesen kann, jedoch wird man in der Regel für Parameterschätzungen Computerprogramme verwenden. Zumindest muß der letzte Parameter signifikant sein und innerhalb der Stationaritäts- und Invertierbarkeitsgrenzen liegen. Die Grenzen sind für den autoregressiven Prozeß der Ordnung  $p$  eingehalten, wenn die Wurzeln  $B_j$  des charakteristischen Polynoms

$$(2.30) \quad (1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p) = 0$$

alle außerhalb des Einheitskreises in der komplexen Zahlenebene liegen. So verstößt ein ARIMA (1,0,0) mit  $\varphi_1 = 2.0$  gegen diese Bestimmung, weil das Polynom



$$(2.31) \quad (1 - 2.0B) = 0$$

nur die Lösung  $B = 1/2$  besitzt: der Prozeß ist nicht stationär.

Ähnliches gilt für die Invertierbarkeitsbedingung. Ein Moving-average-prozeß der Ordnung  $q$  ist invertierbar, wenn die Wurzeln  $B$  des charakteristischen Polynoms

$$(2.32) \quad (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) = 0$$

alle außerhalb des Einheitskreises in der komplexen Zahlenebene liegen.

Danach müssen die Residuen  $a_t$  des vorläufigen Modells zwei Tests bestehen:

a) die ACF der Residuen darf keine signifikanten Werte für die ersten beiden Lags ( $k=1$  und  $k=2$ ) aufweisen:  $H_0: ACF(1)=ACF(2)=0$  und (b) die  $a_t$  müssen sich wie weißes Rauschen verhalten. Ein Test, ob die gesamte ACF nichtsignifikant von der ACF eines Prozesses mit weißem Rauschen abweicht, wurde von Box & Pierce (1970) vorgelegt:

$$(2.33) \quad Q = T \sum_{j=1}^K r_j^2 \quad Q \sim \chi^2_{df=k-p-q}$$

Ist  $Q$  signifikant, muß die Nullhypothese verworfen werden. Die  $\hat{a}_t$  repräsentieren dann kein weißes Rauschen mehr. Verschiedene Autoren empfehlen für  $k$  ca. 10-30 zu wählen, da  $Q$  sehr sensitiv von  $k$  abhängt (s.a. Box & Jenkins, 1976, S. 291).

Weitere Prüfungen des vorläufigen Modells schließen die Berechnung des multiplen  $R^2$  als Maß für die Datenanpassung ein:

$$(2.34) \quad R^2 = 1 - \sum_{t=1}^T (\hat{a}_t^2 / y_t^2) \text{ mit } y_t = \begin{cases} W_t - \mu_w & \text{bei nichtstationärem} \\ \text{Prozeß} \\ Y_t - L & \text{bei stationärem Prozeß} \end{cases}$$

Eine weitere Möglichkeit der Modellprüfung bietet die systematische Aufblähung und Reduktion des Modells. Im überparametrisierten Modell müssen die überflüssigen Parameter nichtsignifikant sein. Im reduzierten (unterparametrisierten) Modell dürfen die Residuen kein weißes Rauschen sein.

Weitere praktische Hinweise zur Identifikation von ARIMA( $p,d,q$ ) Prozessen finden sich bei Glass, Willson & Gottman (1975), Anderson (1975), Gottman & Glas (1978), Makridakis & Wheelwright (1978a,b), Revenstorf & Keeser (1979), Revenstorf (1979), McCain & McCleary (1979) und Gottman (1981).

Zur Veranschaulichung einiger komplizierterer ARIMA (p,d,q)-Modelle wollen wir noch einige Beispiele geben. In Fig. 2.13 haben wir einen nichtstationären IMA-Prozeß vorliegen. Die Nichtstationarität geht auf stochastische Drift zurück, weil  $\theta_0$  nichtsignifikant ist. Die Prozeßgleichung lautet (McCain & McCleary, 1979)

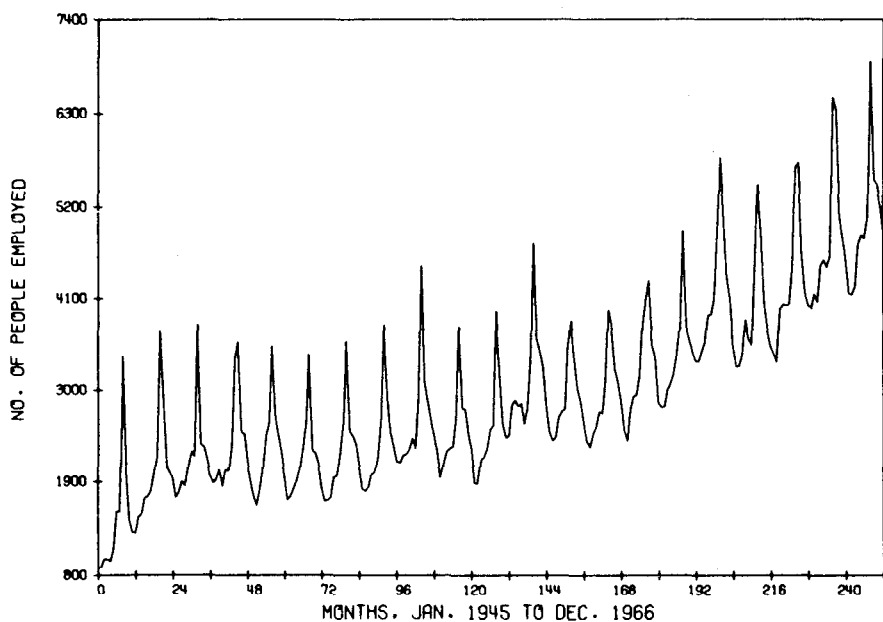


Fig. 2.13: Beispiel für einen ARIMA (0,1,1) (0,1,1),,-Prozeß

$$(2.35a) \quad (1 - B)(1 - B^{12}) Y_t = (1 - .60B)(1 - .68B^{12}) a_t$$

oder

$$(2.35b) \quad Y_t = \frac{(1 - .60B)(1 - .68B^{12})}{(1 - B)(1 - B^{12})} a_t$$

Ein Beispiel für einen in der Varianz nichtstationären Prozeß finden wir in Fig. 2.14:

Die Prozeßgleichung lautet:

$$(2.36) \quad \ln(Y_t) = \frac{(1 - .4321B)(1 - .1884B^{12})}{(1 - B)} a_t$$

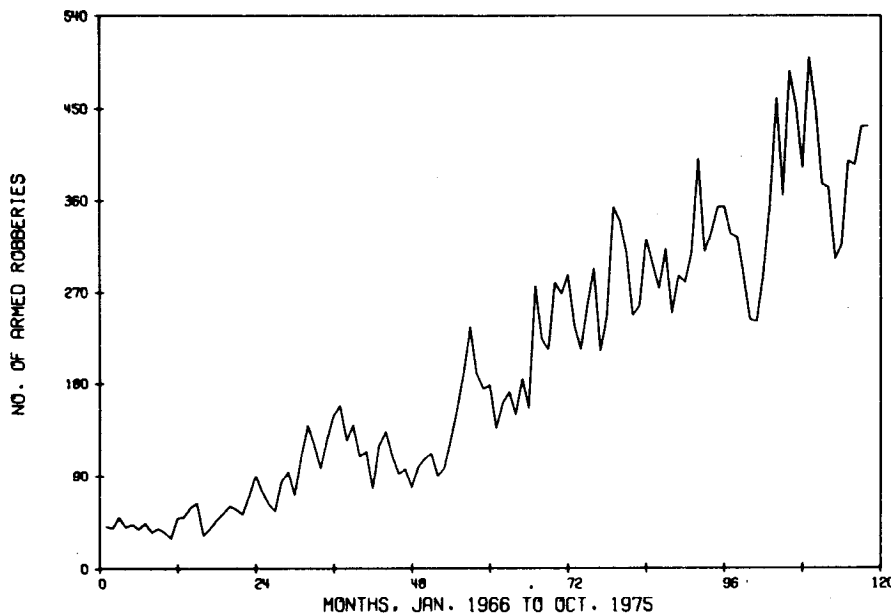


Fig. 2.14: Beispiel für einen ARIMA  $(0,1,1)(0,0,1)_{12}$  für  $\ln(Y_t)$  ( $Y_t$  = Zahl der bewaffneten überfälle in Boston 1966-75)

## 2.8 Multiple Zeitreihenanalyse : Transferfunktionsmodelle

Unter multipler Zeitreihenanalyse verstehen wir die Analyse eines Bedingungs-zusammenhangs einer abhängigen und mehrerer unabhängigen Variablen. Das Modell dieses Zusammenhangs ist die Transferfunktion. Die von Box & Jenkins (1976) vorgeschlagenen Transferfunktionsmodelle können als dynamische Form der multiplen Regression verstanden werden. Wie leicht einzusehen ist, sind Fragestellungen, in denen die Beziehung zwischen einer unabhängigen und mehrerer unabhängiger Variablen untersucht werden soll, nicht mit univariaten Analysen zu beantworten. Multiple Variablenzusammenhänge können z.B. bei folgenden Fragestellungen auftreten: In welcher Weise (a) wirkt sich das Wetter auf das Befinden von Personen aus? (b) wirken sich ökonomische Variable (z.B. Arbeitslosigkeit) auf Verhalten (z.B. Suizidneigung) (Vigderhous, 1978) aus? (c) beeinflussen Tokens das Verhalten eines Probanden im Verlauf einer Therapie (d) verändern neue Telefongebühren die „Telefoniergewohnheiten“ (McSweeny, 1978)?

Allen Beispielen ist trotz inhaltlicher Verschiedenheiten die Heraushebung einer unabhängigen und einer abhängigen Variablen gemeinsam. Die hypothe-

tische Beeinflussung der abhängigen Variablen im Verlauf der Zeit läßt sich am besten graphisch darstellen (s. Figur 2.15).

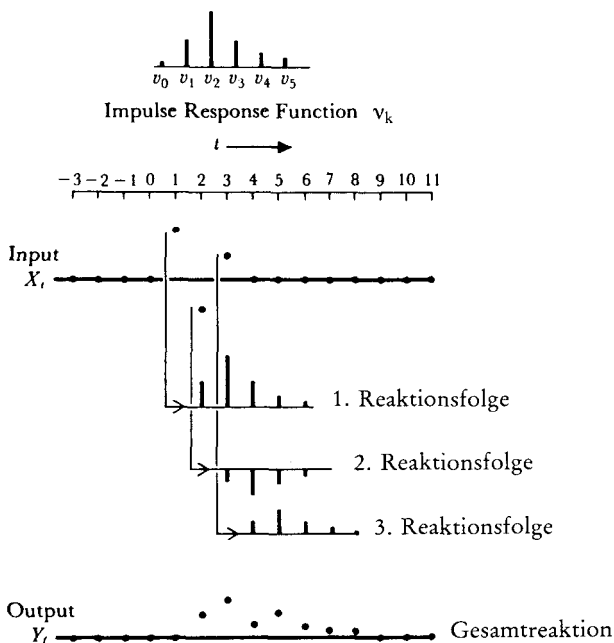


Fig. (2.15): Lineare Beeinflussung (Transfer) der abhängigen Variablen  $Y_t$  durch die unabhängige Variable  $X_t$  im Verlauf der Zeit  $t = -\infty, \dots, +\infty$

Greifen wir zur Erklärung von Figur (2.15) das inhaltliche Beispiel c) heraus. Der Input  $X_t$  sei charakterisiert durch die Tokengabe (zum Zeitpunkt  $t=1$  Übergabe von zwei Tokens, zum Zeitpunkt  $t=2$  Wegnahme von einem Token, zum Zeitpunkt  $t=3$  wiederum Übergabe von einem Token) und die abhängige Variable  $Y_t$  durch das Verhalten der Person. Aus diesen zwei Zeitreihen möchte man auf eine eventuell vorliegende Beziehung von  $X$  auf  $Y$  schließen. Ist  $X$  ein führender Indikator von  $Y$ ? Kann man diese Frage nicht verneinen, kann man untersuchen, in welcher Weise sich der Interventionseffekt über die Zeit hinweg auf  $Y$  verteilt. Diese zeitliche „Verschmierung“ des Interventionseffektes wird durch die Impuls-Response-Funktion  $v_k$  geleistet, wobei sich theoretisch die resultierende Reaktion  $Y_t$  aus einzelnen sich überlagernden Reaktionen zusammensetzt. So würde man sich bei Vorgabe von zwei Tokens zum einzigen Zeitpunkt  $t=1$  die erste Reaktionsfolge mit Maximum zu  $t=3$  denken. Würde allein ein Token zu  $t=2$  weggenommen, wäre nach dem Modell die zweite Reaktionsfolge mit negativem Maximum bei  $t=4$  zu erwarten. Hätte man dagegen zu  $t=3$  ein Token gegeben, würde nach dem Modell

die dritte Reaktionsfolge mit Maximum bei  $t=5$  eintreffen. Die drei Reaktionsfolgen auf die drei Einzeleingriffe setzen sich additiv zur Gesamtreaktionsfolge  $Y_t$  zusammen, wenn statt der isolierten Einzelhandlungen eine Interventionsfolge  $X_t$  auf der Therapeutenseite ablaufen würde. Zur Überprüfung der hypothetischen Modellannahmen dieses zugegebenermaßen sehr einfachen und nicht auf Interaktion ausgerichteten Modells benötigt man die Transferfunktionsanalyse.

Unter einer *bivariaten* Transferfunktion versteht man ein Modell, das die zeitlich verschobenen Einflüsse der exogenen Variablen  $X_t$  auf die abhängige Variable  $Y_t$  beschreibt:

$$(2.37) \quad Y_t = f(X_t) + N_t = Y_t^* + N_t$$

wobei:  $Y_t^*$  den Teil von  $Y_t$  enthält, der exakt durch die Variable  $X$  vorhergesagt bzw. erzwungen werden kann und  $N_t$  Prognosefehler und andere ausgelassene exogene Variable  $U, V, W, Z$  repräsentiert. Dabei folgt  $N_t$  einem  $ARIMA(p, d, q)$   $(P, D, Q)_s$ -Modell

(2.37) ähnelt dem Interventionsmodell in Kap. 3.1.3. Es gibt aber hierzu einen fundamentalen Unterschied. Da die Variable  $X$  nicht gesetzt, sondern beobachtet wird, muß das Transfermodell (2.37) empirisch identifiziert werden, bevor die Parameter geschätzt werden können. Es wird sich herausstellen, daß dieser Vorgang wesentlich schwieriger ist als die Aufstellung eines Interventionsmodells.

Die Beeinflussung von  $Y_t^*$  durch  $X_t$  läßt sich allgemein formulieren als Transfermodell ohne  $N_t$ -Komponente:

$$(2.38a) \quad Y_t^* - \delta_1 Y_{t-1}^* - \dots - \delta_r Y_{t-r}^* = \omega_0 X_{t-b} - \omega_1 X_{t-b-1} - \dots - \omega_s X_{t-b-s}$$

oder in Operatorschreibweise

$$(2.38b) \quad (1 - \delta_1 B - \dots - \delta_r B^r) Y_t^* = (\omega_0 - \omega_1 B - \dots - \omega_s B^s) X_{t-b}$$

oder kürzer

$$(2.38c) \quad \delta(B) Y_t^* = \omega(B) X_{t-b} = \omega(B) B^b X_t = \Omega(B) X_t$$

bzw.

$$(2.38d) \quad Y_t^* = \frac{\Omega(B)}{\delta(B)} X_t = \delta(B)^{-1} \Omega(B) X_t = v(B) X_t$$

Der lineare Filter (2.38d) kann ausführlich geschrieben werden:

$$(2.39) \quad Y_t^* = (v_0 + v_1 B + v_2 B^2 + \dots) X_t = v_0 X_t + v_1 X_{t-1} + v_2 X_{t-2} + \dots$$

Er enthält einen „moving average“ Operator  $w(B)$ , einen „autoregressiven“ Operator  $\delta(B)$  und einen Verzögerungsindex  $b$ , der angibt, um wieviel Zeiteinheiten später eine Änderung von  $X$  eine Änderung von  $Y$  initiiert. Die Indices  $r, s$  spiegeln das „Gedächtnis“ des Transfermodells wieder: „Soweit reicht die Beeinflussung der gegenwärtigen  $Y_t^*$  durch vergangene  $Y^*$ - und  $X$ -Werte. Die Formeln (2.38a-c) sind der Struktur- und die Formel (2.39) der reduzierten Form in Strukturgleichsmodellen (z. B. LISREL) vergleichbar. Der Filter (2.39) repräsentiert die Beziehung zwischen Input und Output, ohne die innere Struktur  $\Omega(B)$  und  $\delta(B)$  zu berücksichtigen. Sie ähnelt der multiplen Regression, wird aber „linearer Filter“ genannt, um den Bezug zur Zeitreihe deutlich zu machen. Das ursprüngliche Modell (2.37) läßt sich dann formulieren als

$$(2.40) \quad Y_t = \frac{\Omega(B)}{\delta(B)} X_t + N_t = \delta(B)^{-1} \Omega(B) X_t + N_t = v(B) X_t + N_t$$

und graphisch darstellen in Figur 2.16 und 2.17)

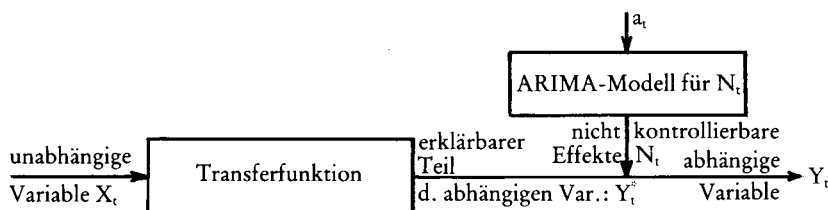


Fig. 2.16: Prozeßmodell des Transfermodells (dynamische Form der multiplen Regression)

Sollte man aus inhaltlichen Gründen einen  $d_x$ -fach differenzierten Input mit einem  $d_y$ -fach differenzierten Output  $Y_t$  verknüpfen wollen, führt das zu:

$$(2.41) \quad (1-B)^{d_y} Y_t = v(B) (1-B)^{d_x} X_t + N_t$$

So sind z.B. bei der Fragestellung: „Werden die Veränderungen der Stimmung  $Y_t$  des Pb Z durch Veränderungen des Wetters  $X_t$  beeinflusst?“ Die Grade der Differenzenbildung sind  $d_x = d_y = 1$  und das Transfermodell nimmt dann die Form  $Y_t = Y_{t-1} + v_0(X_t - X_{t-1}) + N_t$  an.

Der Anteil  $N_t$  an  $Y_t$ , der durch  $X$  nicht erklärt wird, ist ein autokorrelierter Prozeß und daher kein weißes Rauschen  $a_t$ . In  $N_t$  sind neben Meßfehlern auch die Einflüsse ausgelassener Variabler enthalten, so daß  $N_t$  einem meist nicht-stationärem ARIMA(p,d,q) (P,D,Q)-Modell folgt.

Wir sehen jetzt, warum die Frage nach der Beeinflussung von  $Y_t$  durch  $X$  nicht mit der multiplen Regression beantwortet werden darf: (a) die Filterkoeffizienten  $v_k$  können nicht mit der normalen OLS-Regression geschätzt werden, weil die  $N_t$  sich nicht unabhängig verteilen und (b) die Filterkoeffizienten spiegeln nicht die direkten Effekte der verschiedenen  $X$  auf  $Y_t$  wieder. Die direkten Effekte sind nur durch die  $w$ -Parameter repräsentiert. Die Frage, die durch die  $w$ -Parameter von  $Q(B)$  beantwortet werden soll, lautet: „Gibt es in  $X_{t-i}$  direkte zusätzliche Effekte auf  $Y_t$ , die nicht schon in der Geschichte von  $Y_t$  enthalten sind, wobei die Geschichte von  $Y_t$  ja unter ständigem Einfluß der Zeitreihe  $X$  stand (s. Figur 2.17).

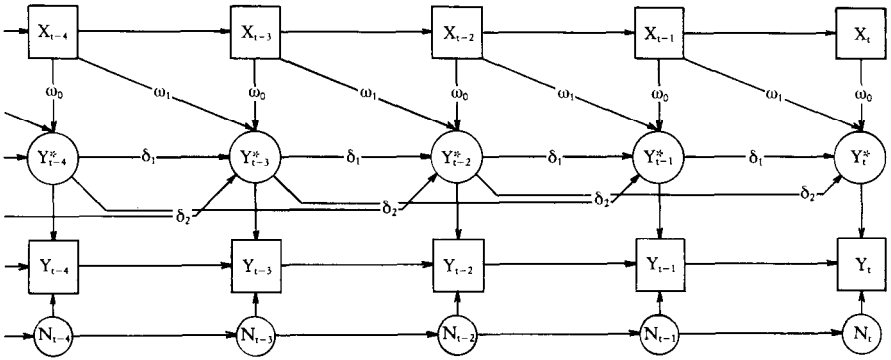


Fig. 2.17: Pfadmodell des bivariaten Transfermodells

$$(1 - \delta_1 B - \delta_2 B^2) Y_t = (\omega_0 - \omega_1 B) X_t + N_t$$

Zur Vervollständigung des Modells (2.40, 2.41) muß die Komponente  $N_t$  nach (2.18) oder (2.19) als ARIMA-Modell abgebildet werden:

$$(2.42a) \quad W_t - \mu_W = \frac{\theta(B)}{\varphi(B)} a_t \quad \text{mit} \begin{cases} W_t = (1-B)^{dN} N_t \\ \mu_W = E(W_t) = \text{Parameter, der} \\ \text{den deterministischen Trend} \\ \text{in } N_t \text{ bestimmt} \end{cases}$$

bzw.

$$(2.42b) \quad N_t = \frac{\mu_W}{(1-B)^{dN}} + \frac{\theta(B)}{(1-B)^{dN} \varphi(B)} a_t$$

Aus (2.40-2.42) läßt sich dann das vollständige Transfermodell mit Veränderungen auf  $X$  und  $Y$  und eventuell nichtstationärem  $N_t$  (s.a. Jenkins, 1979, S. 103) bilden:

$$(2.43a) \quad (1-B)^{dy} Y_t = \frac{\Omega(B)}{\delta(B)} (1-B)^{dx} X_t + \frac{\mu_w}{(1-B)^{dN}} + \frac{\theta(B)}{(1-B)^{dN} \varphi(B)} a_t$$

bzw.

$$(2.43b) \quad (1-B)^{dy+dN} Y_t - \mu_w = \frac{\Omega(B)}{\delta(B)} (1-B)^{dx+dN} X_t + \frac{\theta(B)}{\varphi(B)} a_t$$

Die Filterkoeffizienten  $v_j$  können, wie in Fig. 2.18 gezeigt wird, sehr stark mit den „Gedächtnisindices“  $r, s, b$  variieren.

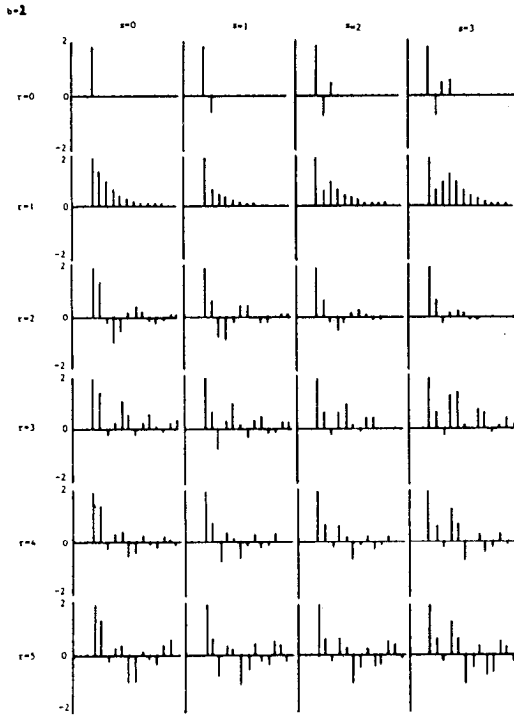


Fig. 2.18: Impuls Response-Parameter  $v_k$  für  $r = 0, 1, 2, 3$  und  $b = 2$  für den Parametersatz  $\delta_1 = .7, \delta_2 = .6, \delta_3 = -.4, \delta_5 = -.2, \omega_1 = .7, \omega_2 = -.5, \omega_3 = -.6$  des Transfermodells (2.38)

Im fehlerfreien theoretischen Modell (2.38) kann man bei Kenntnis der Filterkoeffizienten die Strukturparameter  $\delta$  und  $w$  nach bestimmten Regeln zurückrechnen (Box & Jenkins, 1976, Kap. 10.2.2). Bei der praktischen Datenanalyse sind aber neben den Parametern auch noch die Gedächtnisindices  $r$  sowie der Verzögerungsindex  $b$  unbekannt. Diese müssen festgelegt werden, bevor Parameter geschätzt werden können. Dieses Problem ähnelt der Frage, wieviel



Variable in eine multiple Regression aufzunehmen sind. Erleichtert wird die Beantwortung, wenn unkorrelierte Prädiktoren vorliegen. Bei der Zeitreihenanalyse geht man ähnlich vor.

Man transformiert alle Variablen des Transfermodells

$$(2.44) \quad (1-B)^{dy} Y_t = \frac{\Omega(B)}{\delta(B)} (1-B)^{dx} X_t + N_t$$

so daß die Autokorrelationen der unabhängigen Variablen  $(1-B)^{dx} X_t$  verschwinden. Man partialisiert also alle systematischen Anteile, die nur durch die „Geschichte“ in  $X$  erklärbar und somit redundant sind, heraus. Dieser Vorgang wird „Vorweißen“ genannt. Der vorgeweißte Input  $\alpha_t$  enthält nur noch „neue“ nicht vorhersagbare Information.

Damit die Gleichung (2.44) weiterhin gilt, müssen die Zeitreihen  $(1-B)^{dy} Y_t$  und  $N_t$  derselben Transformation unterworfen werden. Als Ergebnis erhalten wir den in seinen Variablen transformierten Filter:

$$(2.45) \quad \beta_t = v(B)\alpha_t + \varepsilon_t$$

Die neuen Variablenbeziehungen sind graphisch in Figur 2.19 dargestellt. Ein Vergleich mit Figur 2.17 zeigt den neuen - nicht mehr autokorrelierten - Input  $\alpha_t$ :

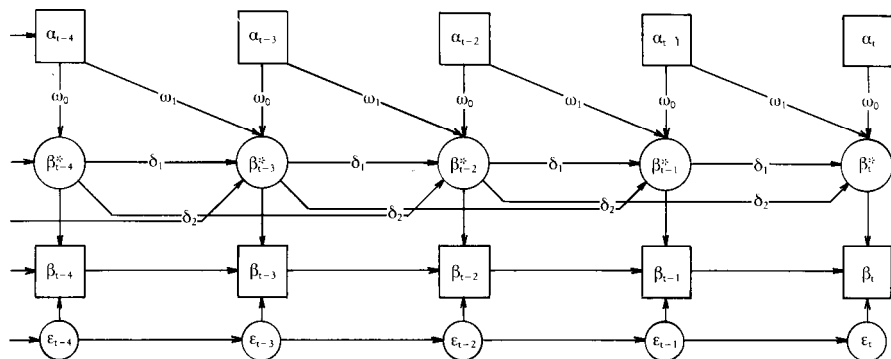


Fig. 2.19: Vorgeweißtes Transfermodell  $\beta_t = v(B)\alpha_t + \varepsilon_t$  mit nicht autokorreliertem Input  $\alpha_t$

Für die Identifikation benötigt man die Kreuzkorrelationsfunktion  $CCF(k) = \varrho_{XY}(k)$  und die Kreuzkovarianzfunktion  $\gamma_{XY}(k)$ .

Bei einem bivariaten stationären Prozeß können wir eine Reihe von Kovarianzen betrachten (Figur 2.20)

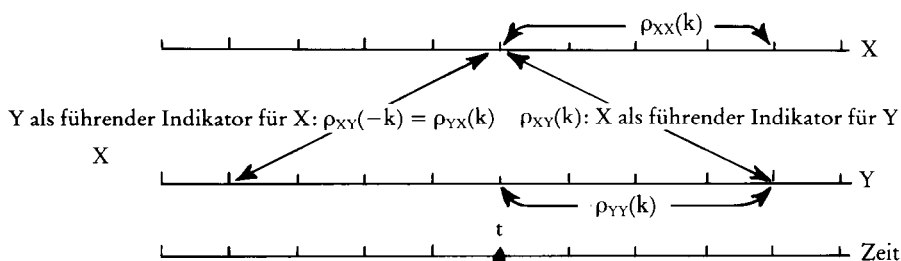


Fig. 2.20: Autokorrelationen und Kreuzkorrelationen eines bivariaten stochastischen Prozesses

Die Kreuzkovarianzen werden für stationäre Zeitreihen  $X_t$ ,  $Y_t$  definiert nach:

$$(2.46a) \quad \begin{aligned} \gamma_{xy}(k) &= E[(X_t - \mu_x)(Y_{t+k} - \mu_y)] \\ &= E[(X_{t-k} - \mu_x)(Y_t - \mu_y)] \end{aligned} \quad k=0,1,2,\dots$$

und

$$(2.46b) \quad \begin{aligned} \gamma_{yx}(k) &= E[(Y_t - \mu_y)(X_{t+k} - \mu_x)] \\ &= E[(Y_{t-k} - \mu_y)(X_t - \mu_x)] \end{aligned} \quad k=0,1,2,\dots$$

Die Korrelationsfunktion ergibt sich durch Division von (2.46) durch die Streuungen:

$$(2.47) \quad \rho_{xy}(k) = \gamma_{xy}(k) / (\sigma_x \sigma_y) \quad k=0, \pm 1, \pm 2, \dots$$

Für die Identifikation des Transfermodells (2.38) ist wichtig, daß die Kreuzkorrelationen zwischen den vorgeweißten Prozessen für bestimmte Lags  $k$  gleich Null sind.

Man kann die Kreuzkorrelationen schätzen mit:

$$(2.48) \quad \text{cov}_{xy}(k) = \begin{cases} \frac{1}{T} \sum_{t=1}^{T-k} (X_t - \bar{X})(Y_{t+k} - \bar{Y}) & k = 0, 1, 2, \dots \\ \frac{1}{T} \sum_{t=1}^{T+k} (Y_t - \bar{Y})(X_{t-k} - \bar{X}) & k = 0, -1, -2, \dots \end{cases}$$

$$(2.49) \quad r_{xy}(k) = \text{cov}_{xy}(k) / (\sqrt{\text{cov}_{xx}(0)} \sqrt{\text{cov}_{yy}(0)}) \quad k = 0, \pm 1, \pm 2$$

Ein Beispiel für eine Anwendung der Kreuzkorrelation im Bereich der klinischen Psychologie findet sich bei Revenstorf & Keeser (1978) (s.a. Figur 2.21)

in einer Studie zu Steuerungsmöglichkeiten des Zigarettenkonsums. Von Interesse ist hier die Frage, ob die Motivation ein führender Indikator für Zigarettenkonsum in einer Therapie ist.

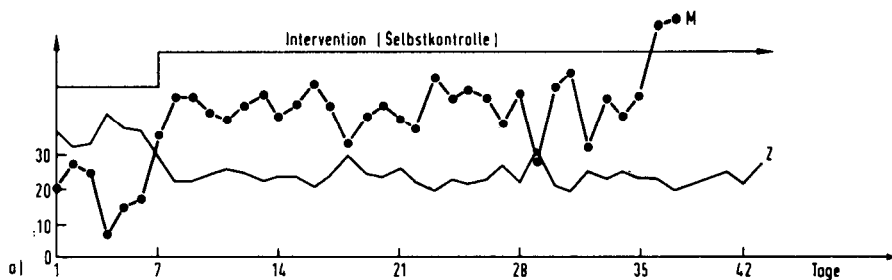


Fig. 2.21a: Zigarettenkonsum (Z) und Therapiemotivation (M) eines Rauchers an 42 aufeinander folgenden Tagen

Die beiden neuen durch Vorweißen gewonnenen Zeitreihen  $\alpha_t$  und  $\beta_t$  können jetzt zur Identifikation von  $r$ ,  $s$  und  $b$  herangezogen werden, da die Filterkoeffizienten  $v_k$  jetzt einfache mit Standardabweichungen verzerrte Kreuzkorrelationen sind

$$(2.50) \quad v_k = Q_{\alpha\beta}(k) \frac{\sigma_\beta}{\sigma_\alpha}$$

Eine ähnliche Beziehung ist aus der Regression bekannt.

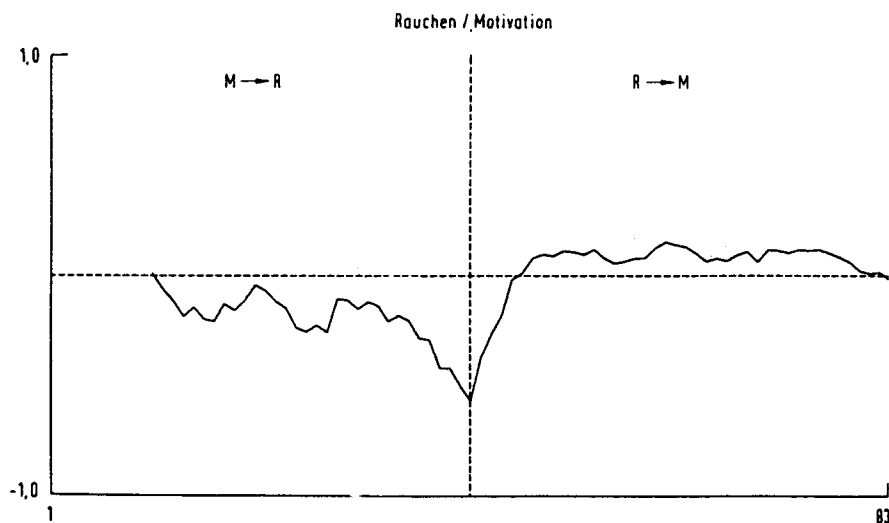


Fig. 2.21b: Kreuzkorrelationsfunktion von Zigarettenkonsum (Z) und Therapiemotivation (die vom Betrag größte Korrelation ist die synchrone Korrelation mit Lag  $k=0$ )

Dort ist das Regressionsgewicht ebenfalls eine mit den Maßstabstreuungen verzerrte Korrelation. Eine Schätzung für (2.50) ergibt sich nach

$$(2.51) \quad \hat{v}_k = r_{\alpha\beta}(k) \frac{s_\beta}{s_\alpha} \quad k=1,2,3,\dots$$

Hätte man die  $v_k$  mit den Originalzeitreihen schätzen wollen, wäre die Analogie zur Regression verloren gegangen, weil in die Schätzungen noch die Autokorrelation  $r_{,,}(k)$  der unabhängigen Variablen eingegangen wäre (s. a. McCleary & Hay, 1980, S. 248). Die Autokorrelationen der Zeitreihen treiben die Kreuzkorrelationen von X und Y hoch und suggerieren einen engen Zusammenhang.

Für die Kreuzkorrelationen zweier Variablen, von der die eine nur weißes Rauschen enthält (z.B.  $\alpha_t$ ), hat Bartlett (1955) eine Approximation für das 95%ige Konfidenzintervall

$$(2.52) \quad r_{\alpha\beta}(k) - \frac{2.0}{\sqrt{T-k}} < \varrho_{\alpha\beta}(k) < r_{\alpha\beta}(k) + \frac{2.0}{\sqrt{T-k}}$$

angegeben.

Sind die  $\hat{v}_k$  jetzt bekannt, können die „Gedächtnis“-indices r und s und der Verzögerungsindex b vorläufig bestimmt werden. Hinweise hierfür finden sich bei Box & Jenkins (1976<sup>3</sup>, S. 347).

## Schätzung und Modelltests

Schätzung und Modelltests verlaufen wieder nach einem iterativen Schema (Schätzung evtl. mit dem Computerprogramm von Pack, 1978):

1. Jede nichtstationäre Zeitreihe bis zur Stationarität differenziert
- 2. Das ARIMA-Modell für die unabhängige Variable X (2.44) wird invertiert (2.45) und der invertierte Operator wird auf alle Variablen des Transfer-N-Modells angewendet (Vorweißen) (2.45)
- 3. Die Kreuzkorrelationsfunktion der vorgeweißten Zeitreihen  $\alpha_t$ ,  $\beta_t$  dient zur Identifikation des Transfermodells (s. Box & Jenkins, 1976<sup>3</sup>, S. 347). Die Residuen dieses Modells dienen zur Identifikation des  $N_t$ -Modells.
- 4. Schätzung aller Parameter des Transfermodells und des ARIMA-Modells des  $N_t$ -Prozesses (alle Parameter müssen signifikant und in den erlaubten Grenzen der Stabilität, Invertierbarkeit, Stationarität liegen)
- 5. Residuen des gesamten Modells müssen weißem Rauschen entsprechen und dürfen nicht mit dem Regressor  $X_t$  korrelieren (gilt für alle lags k): d.h. alle Kreuzkorrelationen zwischen den Residuen und X müssen nichtsignifikant sein.
6. Modellinterpretation

### *Ein empirisches Beispiel*

In der Literatur sind bisher noch nicht viele Anwendungen der Transfermodelltechnik veröffentlicht worden. Ausnahmen sind Helmer & Johansson (1977) und Vigderhous (1978). Die Ergebnisse der letzten Arbeit sollen kurz referiert werden.

Ausgehend von inhaltlichen Überlegungen von Durkheims Theorie zum Selbstmord (1897) wollte Vigderhous den Einfluß der Arbeitslosenquote auf die Suizidneigung (hier: Suizidrate) in den USA untersuchen. Die Zeitreihe findet sich in Figur (2.22)

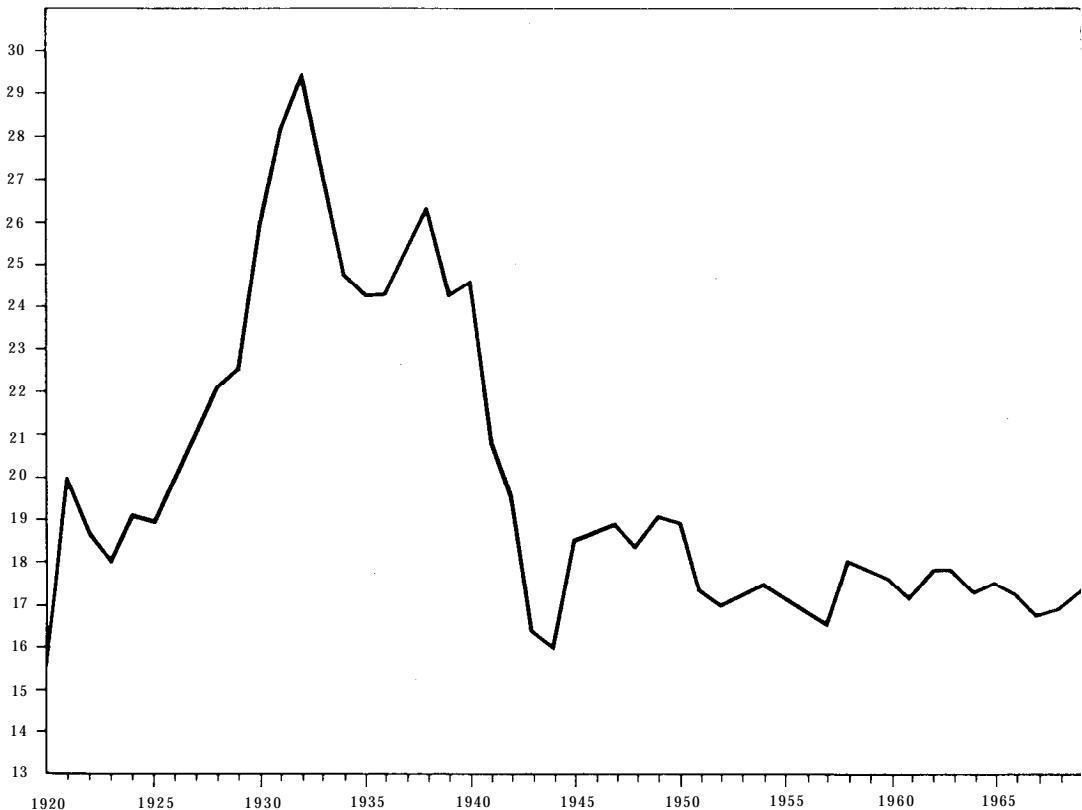


Fig. 2.22: Suizidraten pro 100000 Weiße männlichen Geschlechts zwischen 1920-1969

Das saisonale  $ARIMA(2,0,0)(0,0,1)_{10}$ -Modell für die Zeitreihe  $X_t$  der unabhängigen Variablen „Arbeitslosenquote“ war:

$$(2.53) \quad X_t - \mu_x = \frac{(1 - \theta_{10} B^{10})}{(1 - \varphi_1 B - \varphi_2 B^2)} a_t = \frac{\theta(B)}{\varphi(B)} a_t \quad \text{mit} \quad \begin{aligned} \hat{\mu}_x &= 6.54 \\ \hat{\theta}_{10} &= 0.38 \\ \hat{\varphi}_1 &= 1.16 \\ \hat{\varphi}_2 &= 0.29 \end{aligned}$$

Auffallend ist die moving-average-Komponente, die auf einen Zehnjahresrhythmus in der Arbeitslosenquote hindeutet. Mit Hilfe von (2.53) läßt sich  $X_t - \mu_x$  zu weißem Rauschen transformieren

$$(2.54) \quad \frac{\varphi(B)}{\theta(B)} (X_t - \mu_x) = \alpha_t = a_t$$

Dieselbe Transformation  $\varphi(B)/\theta(B)$  wird auf die abhängige Zeitreihe  $Y_t$  der Suicidraten und auf  $N_t$  angewendet. Man erhält dann die vorgeweißten Zeitreihen  $\beta_t$  und  $\varepsilon_t$  aus (2.45).

Die Schätzung der Impuls-Response-Parameter (2.51) des Filters (2.39 bzw. 2.45)

$$\hat{v}_0 = .36, \hat{v}_1 = -.007, \hat{v}_2 = .042, \hat{v}_3 = -.112 \text{ mit Index } b=0$$

deuten auf die Synchronität beider Zeitreihen hin. Entweder kovariieren beide ohne Zeitverzug (Lag = 0) (hängen also von einer gemeinsamen Drittvariablen ab), oder die Arbeitslosenquote übt einen sofortigen Einfluß auf die Suizidrate aus.

Das Transfermodell (2.43) wurde als

$$Y_t + 2.296 = 0.3529 X_t + \frac{1}{(1 - 0.84B)} a_t$$

geschätzt (Vigderhous, 1978, S. 44). Alle Parameter sind auf 5% signifikant.

## 2.9 Multivariate Zeitreihenanalyse

Die *univariate Zeitreihenanalyse* ist für Psychologen meist nur im Rahmen des Transfermodells (2.37) zur Abbildung der  $N_t$ -Komponente wichtig. Ähnlich liegen die Verhältnisse zwischen multivariater Zeitreihenanalyse und dem multivariaten Transfermodell. Das multivariate Transfermodell kann dabei als dynamische Erweiterung der multivariaten Regression angesehen werden, wobei die nicht durch unabhängige Variable kontrollierbaren Anteile der abhängigen Zeitreihen einem multivariaten stochastischen Prozeß folgen.

Nehmen wir z.B. die Interaktion einer Dyade. Die Zeitreihe  $Y_{1t}$  repräsentiert z.B. die Häufigkeit einer Verhaltenskategorie A zum Zeitpunkt t bei Person 1

und  $Y_{2t}$  die Häufigkeit einer Verhaltenskategorie B bei Person 2. Damit die geschätzte Beziehungsstruktur zwischen den beiden Zeitreihen von ihren jeweiligen Autokorrelationen nicht verfälscht wird, müssen die redundanten Teile der Zeitreihen auspartialisiert (d.h. „vorgeweißt“) werden. Folgen z.B.  $Y_{1t}$  und  $Y_{2t}$  den univariaten ARIMA-Modellen (2.18, 2.19):

$$(2.55a) \quad (1 - B)Y_{1t} = \theta_{10} + \alpha_{1t}$$

$$(2.55b) \quad (1 - B)Y_{2t} = \theta_{20} + (1 - \theta_{11}B)\alpha_{2t}$$

Analog zu der univariaten Zeitreihenanalyse wird jetzt die Beziehungsstruktur der beiden vorgeweißten Zeitreihen  $\alpha_{1t}$  und  $\alpha_{2t}$  mittels der Kreuzkorrelationsfunktion CCF(k) und der partiellen Kreuzkorrelationsfunktion PCCF(k) untersucht. Moving-average-Prozesse werden durch die CCF(k) und Kreuzregressive Komponenten durch die PCCF(k) identifiziert. Nehmen wir für unser Beispiel an, daß die PCCF(k) für alle k nichtsignifikant ist. Die CCF(k) soll den Verlauf von Figur 2.23 haben. Nur die zwei Kreuzkorrelationen  $r_{\alpha_1\alpha_2}$  (+ 1) und  $r_{\alpha_1\alpha_2}$  (- 1) fallen aus dem 95%igen Konfidenzintervall (2.52) und sind somit signifikant von Null verschieden.

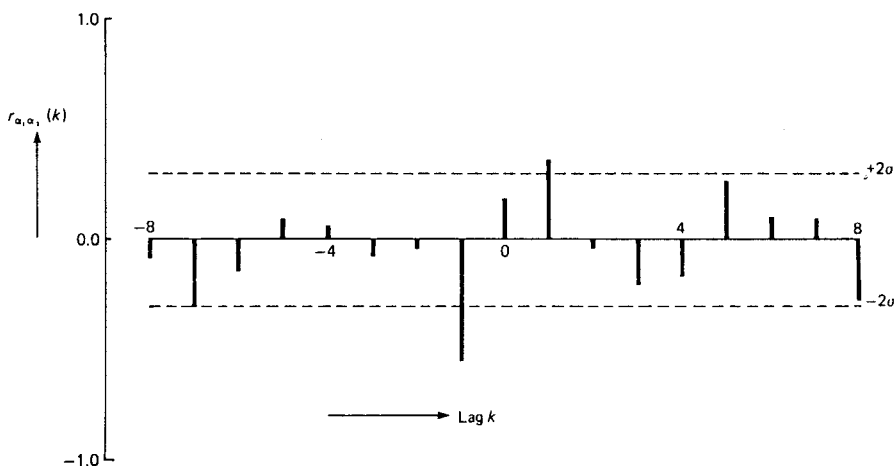


Fig. 2.23: Kreuzkorrelationsfunktion zwischen den zwei vorgeweißten Zeitreihen  $\alpha_{1t}$  und  $\alpha_{2t}$

Ein signifikantes  $r_{\alpha_1\alpha_2}$  (+1) und  $r_{\alpha_1\alpha_2}$  (-1) bedeutet bei gleichzeitigem nicht-signifikanten PCCF(k), daß für die  $\alpha_t$  ein bivariater Moving-Average-Prozeß angesetzt werden muß:

$$(2.56a) \quad \alpha_{1t} = a_{1t} - \theta_{12}a_{2,t-1}$$

$$(2.56b) \quad \alpha_{2t} = -\theta_{21}a_{1,t-1} + a_{2t}$$

Setzt man (2.56) in (2.55) ein, erhält man

$$(2.57a) \quad (1-B)Y_{1t} = \theta_{10} + a_{1t} - \theta_{12}Ba_{2t}$$

$$(2.57b) \quad \begin{aligned} (1-B)Y_{2t} &= \theta_{20} + (1-\theta_{11}B)(-\theta_{21}Ba_{1t} + a_{2t}) \\ &= \theta_{20} + (-\theta_{21}B + \theta_{11}\theta_{21}B^2)a_{1t} + (1-\theta_{11}B)a_{2t} \end{aligned}$$

oder allgemein in Operatorschreibweise

$$(2.58a) \quad \varphi_{11}(B)(1-B)Y_{1t} = \theta_{10} + \theta_{11}(B)a_{1t} + \theta_{12}(B)a_{2t}$$

$$(2.58b) \quad \varphi_{22}(B)(1-B)Y_{2t} = \theta_{20} + \theta_{21}(B)a_{1t} + \theta_{22}(B)a_{2t}$$

läßt sich (2.58) als Matrixgleichung schreiben:

$$(2.59) \quad \begin{bmatrix} \varphi_{11}(B) & \varphi_{12}(B) \\ \varphi_{21}(B) & \varphi_{22}(B) \end{bmatrix} \cdot \begin{bmatrix} (1-B)^{d_1}Y_{1t} \\ (1-B)^{d_2}Y_{2t} \end{bmatrix} = \begin{bmatrix} \theta_{10} \\ \theta_{20} \end{bmatrix} + \begin{bmatrix} \theta_{11}(B) & \theta_{12}(B) \\ \theta_{21}(B) & \theta_{22}(B) \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}$$

mit:  $d_1=d_2=1$

und den Operatoren:  $\varphi_{11}(B)=\varphi_{22}(B)=1$

$$\varphi_{12}(B)=\varphi_{21}(B)=0$$

$$\theta_{11}(B)=1, \quad \theta_{12}(B)=-\theta_{12}$$

$$\theta_{21}(B)=(-\theta_{21}B + \theta_{11}\theta_{21}B^2)$$

$$\theta_{22}(B)=(1 - \theta_{11}B)$$

(2.59) läßt sich noch kompakter formulieren als:

$$(2.60) \quad \Phi(B) (1-B)^d Y_t = \Theta_0 + \Theta(B)a_t$$

Dabei ist (2.60) die multivariate Erweiterung zu (2.18b). Definiert man analog zu (2.19) den Prozeß in Abweichungen ( $W_t - \mu_w$ ) (mit  $W'_t = ((1-B)^{d_1}Y_{1t}, \dots, (1-B)^{d_m}Y_{mt})$ ) kann man (2.60) auch als

$$(2.61a) \quad \Phi(B)(W_t - \mu_w) = \Theta(B)a_t \quad \text{multivariates ARIMA(P,d,Q)-Modell}$$

$$(2.61b) \quad (W_t - \mu_w) = \Phi^{-1}(B)\Theta(B)a_t \text{ mit: } \mu_w = \Phi^{-1}(B)\Theta_0$$

schreiben (Jenkins, 1979, S. 111). Die  $a_t$  besitzen dabei die Kovarianzen  $\sigma_{ij} \neq 0$  für Lag  $k=0$  und  $\sigma_{ij} = 0$  für Lag  $k \neq 0$ . Das ARIMA(P,d,Q)-Modell läßt sich ebenfalls wie das univariate um saisonale Komponenten erweitern.

Die Stationaritätsbedingung verändert sich von der univariaten (2.30) zur multivariaten (2.62) und die Invertierbarkeitsbedingung von (2.32) zu (2.63). Die Wurzeln der Determinantengleichungen



$$(2.62) \quad |\varphi(B)| = 0 \quad \text{multivariate Stationaritätsbedingung}$$

$$(2.63) \quad |\Theta(B)| = 0 \quad \text{multivariate Invertierbarkeitsbedingung}$$

müssen außerhalb des Einheitskreises in der komplexen Zahlenebene liegen.

Das Output-Modell eines multivariaten nichtsisonalen ARIMA(P,d,Q)-Modells findet sich in Fig. 2.24

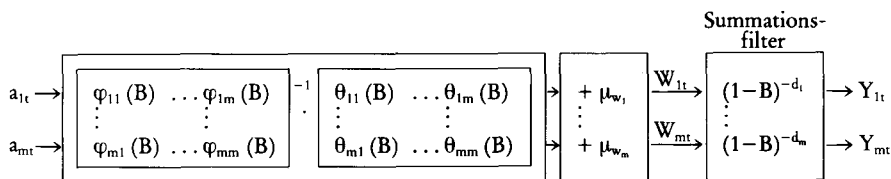


Fig. 2.24: Output-Modell eines multivariaten ARIMA-Modells

Auf andere Modellformen, die z.T. auch unkorrelierte  $a_t$  zu einem Lag  $k=0$  zulassen, verweisen z.B. Granger & Newbold (1976, 1977), Box & Tiao (1977) sowie Jenkins (1979).

## 2.10 Multiple und multivariate Transfermodelle

### 2.10.1 Multiple Transfermodelle

Nach der Konzeption des multiplen Transfermodells beeinflussen mehrere unabhängige Zeitreihen  $X_{jt}$  (=Inputs) eine abhängige Zeitreihe  $Y_t$ :

$$(2.64) \quad Y_t = f_1(X_{1t}) + f_2(X_{2t}) + \dots + N_t = Y_{1t}^* + Y_{2t}^* + \dots + N_t$$

Das Modell (2.64) kann als dynamische Erweiterung der multiplen Regression angesehen werden. Da jedoch die Kreuzkorrelationen autokorrelierter Zeitreihen sehr hoch und die Residuen ebenfalls autokorreliert sind, dürfen die Parameter des multiplen Transfermodells nicht nach der in der multiplen Regression üblichen Methode der kleinsten Quadrate geschätzt werden.

Wir erweitern den Ansatz des einfachen Transfermodells (2.40) entsprechend (2.64) auf mehrere Inputs  $X_{jt}$  ( $j=1, \dots, n$ ):

$$(2.65) \quad Y_t = \frac{\Omega_1(B)}{\delta_1(B)} X_{1t} + \frac{\Omega_2(B)}{\delta_2(B)} X_{2t} + \dots + \frac{\Omega_n(B)}{\delta_n(B)} X_{nt} + N_t$$

Entsprechend (2.41) dürfen die Rohwerte  $Y_t, X_{jt}$  aus *inhaltlichen* Gründen unterschiedlich oft differenziert werden, so daß in (2.65) dann Differenzen statt Rohwerte auftauchen:

$$(2.66) \quad (1 - B)^{d_Y} Y_t = \sum_{j=1}^n \frac{\Omega_j(B)}{\delta_j(B)} (1 - B)^{d_{X_j}} X_{jt} + N_t$$

Anschließend ist noch das eventuell nichtstationäre, saisonale ARIMA-Modell der  $N_t$ -Komponente zu identifizieren und die Parameter zu schätzen. Man erhält dann die multiple Erweiterung von (2.43). Die  $\phi$ -Parameter werden wieder wie in (2.38) als direkte Effekte von  $X_{jt}$  auf  $Y_t$  interpretiert: „... given that the effect of  $X_t$  on  $Y_t$  may already be contained in the past history of  $Y_t$ , is there any additional information contained in  $X_t$ ?“ (Jenkins, 1979, S. 19).

Die Modellidentifikation des multiplen Modells erfolgt in mehreren Schritten. So wird jede  $X_{jt}$ -Zeitreihe zu  $\alpha_{jt}$  vorgeweißt. Derselben Transformation wird dann die eine abhängige Zeitreihe  $Y_t$   $m$ -fach unterworfen. Man erhält dann die Zeitreihen  $\beta_{jt}$ . Aus den Kreuzkorrelationsfunktionen  $\phi_{\alpha\beta j}(k)$  werden die Teilmodelle  $v_j(B) = \frac{\Omega_j(B)}{\delta_j(B)}$  vorläufig identifiziert. Danach müssen (a) ein ebenfalls vorläufiges ARIMA-Modell für  $N_t$  gefunden und (b) Modelltests durchgeführt werden. Als erste Approximation des ARIMA(p,d,q)-Modells für  $N_t$  wählt man oft das ARIMA(p,d,q)-Modell für  $Y_t$ . Natürlich muß dieser Versuch revidiert werden, wenn die Modelltests dieses anschließend fordern. Der Prozeß der Identifikation und Schätzung ist abgeschlossen, wenn das Modell nicht überparametrisiert ist, alle Parameter in den zulässigen Grenzen liegen (gilt auch für die Filterkoeffizienten  $v_j(B)$ : Stabilität des Transfermodells),  $a_t$  weißes Rauschen repräsentiert und minimale Fehlervarianz  $\sigma_a^2$  besitzt.

### 2.10.2 Multivariate Transfermodelle

Im multivariaten Transfermodell (einer dynamischen Erweiterung der multivariaten Regression) beeinflussen mehrere unabhängige Zeitreihen  $X_{jt}$  (= Inputs) mehrere abhängige Zeitreihen  $Y_{it}$  (= Outputs).

Wir setzen jetzt für jedes  $Y_{it}$  ( $i=1, \dots, m$ ) ein multiples Transfermodell (2.66) an und erhalten somit  $m$  Transferfunktionen:

$$(2.67) \quad \begin{aligned} (1 - B)^{d_{Y_i}} Y_{it} &= \sum_{j=1}^n \frac{\Omega_{ij}(B)}{\delta_{ij}(B)} (1 - B)^{d_{X_j}} X_{jt} + N_{it} \\ &= \sum_{j=1}^n v_{ij}(B) (1 - B)^{d_{X_j}} X_{jt} + N_{it} \end{aligned}$$

Saisonale Modelle lassen sich in ähnlicher Weise formulieren.

Die  $m$  Komponenten  $N_{it}$  können zum stochastischen Vektor  $N_t$  zusammengefaßt werden. Er folgt einem eventuell nichtstationären, saisonalen multivariaten ARIMA-Prozeß. Nach mehreren Iterationen des Identifikations-Schätz- und Prüfverfahrens wird man dann das endgültige Modell aufgestellt haben. Das Input-Output-Modell ist schematisch in Fig. 2.25 dargestellt.

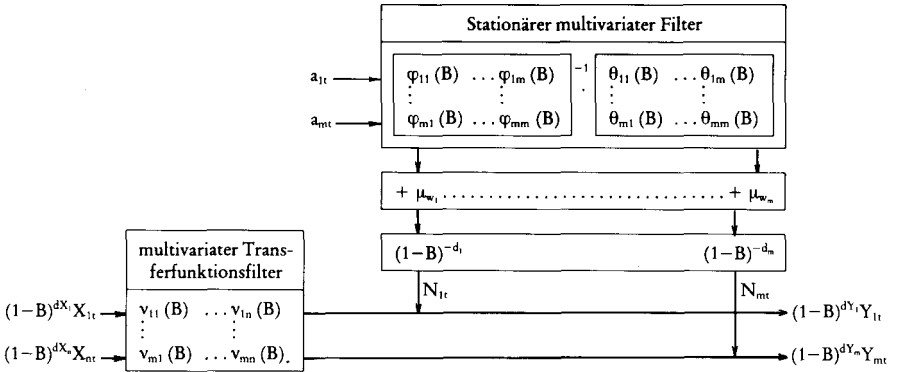


Fig. 2.25: Input-Output-Modell des multivariaten Transfermodells

### 3. Zeitreihenexperimente

$$(T \geq 2, M = 1)$$

Neben den klassischen Experimenten mit randomisierter Zuweisung von Personen auf die experimentellen Versuchsbedingungen haben in letzter Zeit die auf Campbell (1963) und Campbell & Stanley (1963, 1966) zurückgehenden Quasiexperimente an Boden gewonnen. In ihnen gibt es die Kontrolltechnik der Randomisierung nicht. Man unterscheidet zwei Hauptgruppen von Quasiexperimenten: a) das nichtäquivalente Kontrollgruppendesign (s. Cook & Campbell, 1979; Reichardt, 1979) und b) das Zeitreihenexperiment mit einer ganz bestimmten Versuchseinheit. In einem Zeitreihenexperiment werden vor und nach der Intervention längere Zeit Beobachtungen in festen Zeitabständen erhoben. In  $N = 1$  Studien, die auch „intensive designs“ genannt werden, besteht die Untersuchungseinheit meist aus einer nicht zufällig ausgewählten Person (s. McCain & McCleary, 1979). Für  $N > 1$ -Studien hat sich der Begriff „Querschnitt-Zeitreihenexperiment“ („Cross-sectional time series experiment“) eingebürgert (Simonton, 1977).

### 3.1 N = 1-Experimente

Es gibt u.a. zwei gewichtige Gründe für das intensive Design. Zum einen kann man daran interessiert sein, ob man das Verhalten einer ganz bestimmten Person ändern konnte (ob also eine Intervention erfolgreich war). Zum anderen kann man an der Effektivität eines Treatments interessiert sein, zugleich aber befürchten, daß das Treatment nur für einen kleinen Personenkreis anschlägt. In diesem Fall würde eine klassische Untersuchung, bei der ja Effekte über Personen aggregiert werden, u.U. keine signifikanten Ergebnisse bringen. Es ist dann angebracht, das Treatment an jeder einzelnen Person der Stichprobe zu prüfen.

Speziell in der verhaltenstherapeutisch orientierten klinischen Psychologie wurden eine Reihe von N=1 Designs betrachtet (Chassan, 1979<sup>2</sup>; Barlow & Hersen, 1973; Hersen & Barlow, 1976; Kazdin, 1976; Fichter, 1978; Tyler & Brown, 1968), die aber in den meisten Fällen auch bei  $N > 1$ -Untersuchungen hätten Anwendung finden können. Meist begnügt man sich nicht damit, nach einer gewissen Beobachtungszeit ein Treatment einzuführen, sondern man gliedert den Versuchsplan in mehrere A und B Phasen (z. B.  $A_1B_1A_2B_2$ ), wobei  $A_1$  in den sogenannten „base-line“-Designs eine präexperimentelle Beobachtungsphase darstellt. Diese wird dann von einer Treatment- oder B-Phase abgelöst. Dann folgt wieder eine Beobachtungs- oder Lösungsphase  $A_2$  etc. Ein Beispiel für ein  $A_1B_1C_1B_2C_2$  Experiment findet sich in Figur 3.1. Als akzeptablen Kompromiß zwischen Aufwand und Präzision wird von vielen Autoren das  $A_1B_1A_2B_2$ -Design bezeichnet.

Nach einer Beobachtungsphase (Basislinie)  $A_1$  erfolgt die Intervention  $B_1$ . Nach dem Absetzen der Intervention wird wieder eine Basislinie  $A_2$  beobachtet, der dann wieder eine Intervention  $B_2$  folgt. Untersuchungsziel ist es, Wechsel in den Werten der abhängigen Variablen  $Y_t$  nachzuweisen, die mit dem Wechsel in der unabhängigen Variablen (Intervention) korrespondieren. Dabei wird in der klinischen Praxis oft aus ethischen Gründen das Experiment nach der  $A_1B_1$ -Phase abgebrochen, weil der Verlauf der abhängigen Variablen „erfolgsversprechend“ und ein designgesteuertes Absetzen und Wiederaufnehmen des Treatments nicht Verantwortbar erscheint.

Es sind eine Reihe von Interventionseffekten denkbar (s. Figur 3.2), die auf ihre statistische Signifikanz hin überprüft werden können.

Dabei wirft die statistische Auswertung der Zeitreihenexperimente Probleme auf, die in den sonst üblichen Querschnittuntersuchungen mit Zufallsstichproben und randomisierter Zuteilung der Personen auf die experimentellen Bedingungen nicht auftreten.

So ist die externe Validität eines N=1 Zeitreihenexperiments gleich dreifach gefährdet (Levin, Marascuilo & Hubert, 1978): (a) Da die zu untersuchende

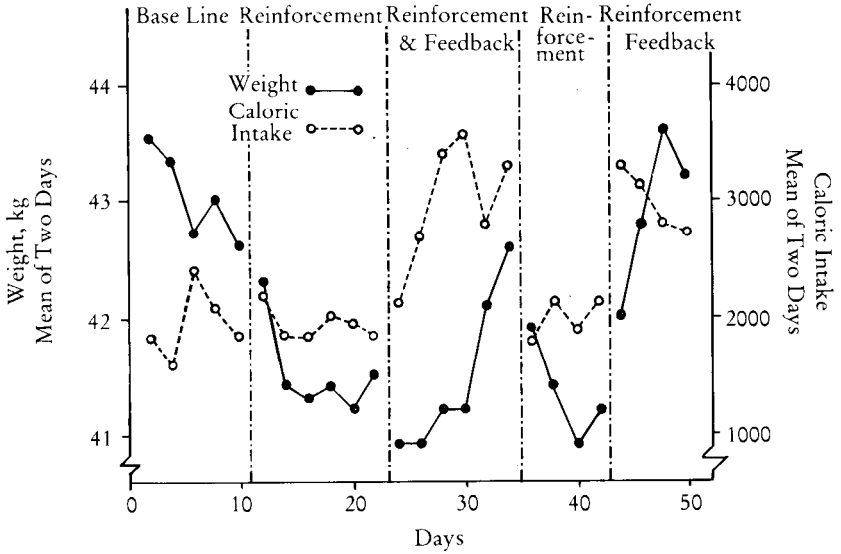


Fig. 3.1: Data from an experiment examining the effect of feedback on the eating behavior of a patient with anorexia nervosa (Patient 4). (Fig. 3, p. 283, from: Agras, W. S., Barlow, D. H., Chapin, H. N., Abel, G. G., and Leitenberg, H. Behavior modification of anorexia nervosa. Archives of General psychiatry, 1974, 30, 279-286. Reproduced by Permission.)

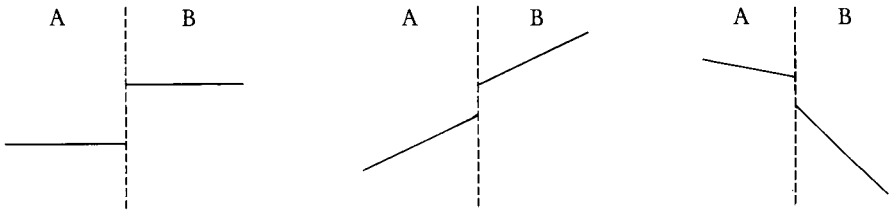


Fig. 3.2: Verlaufsmuster für die  $A_1B_1$ -Phasen eines AB-Designs

Person gezielt und nicht zufällig aus einer Population ausgewählt wurde, ist die Populationsvalidität (Bracht & Glass, 1968) nicht gegeben. Auch wird die Person nicht in zufällig ausgewählten Reiz/Situationskombinationen beobachtet, so daß (b) die ökologische Validität (Bracht & Glass, 1968) beeinträchtigt ist. Zum dritten sind die Beobachtungszeitpunkte oft ebenfalls nicht randomisiert ausgewählt worden, so daß (c) eine Generalisation der Untersuchungsergebnisse auf andere Zeitpunkte unzulässig ist („within-subject generalizability

limitation“). Allerdings sieht Edgington (1980, S. 248) allein in der Strategie der randomisierten Auswahl keine Lösung des Problems: „Thus neither random sampling of a population of subjects nor of a population of treatment times would justify the determination of significance by parametric probability tables for one-subject experimental data.“ (Edgington, 1980, S. 248)

Die interne Validität des  $N=1$  Zeitreihenexperiments wirft ebenfalls Probleme auf, die nur teilweise lösbar sind. So liegt nur dann interne Validität vor, wenn alle äußeren Einflüsse so ausgeschaltet sind, daß die Ergebnisse nur durch experimentelle Manipulation erklärbar sind. Eine unerläßliche Bedingung für ein intern valides Experiment ist dabei die randomisierte Zuteilung von Personen auf experimentelle Bedingungen (Campbell & Stanley, 1966). Sie ist für  $N=1$  Studien unerfüllbar. Es lassen sich aber die A und B Phasen in ihrer Reihenfolge randomisiert vorgeben. So meint Edgington (1980): „... statistical tests whose significance is based on random assignment can be validly applied to one-subject experimental data when there has been random assignment of treatment times to treatments (Edgington, 1967; Revusky, 1967)“ (Edgington, 1980, S. 249).

Ein anderes Problem liegt in der u.U. vorliegenden seriellen Abhängigkeit der Daten bzw. der Residuen (wenn die Auswertung nach dem allgemeinen linearen Modell folgt (s.a. Moosbrugger, 1978). Sie kann durch Ermüdung, Lern- oder andere Carry-Over-Effekte entstehen. Tests, (z.B. t-Test, Varianzanalyse etc.), die von der Unabhängigkeit der Residuen ausgehen, sind in ihren Signifikanzaussagen daher oft irreführend (Hibbs, 1974). Sind die Residuen positiv autokorreliert, sind die berechneten t- oder F-Brüche „zu groß“, d.h.: Interventionen erscheinen als effektiv, während „in Wirklichkeit“ nur ein stochastischer Datentrend vorlag. Eine Reihe von Autoren (Shine & Bower, 1971; Gentile, Roden & Klein, 1972; Hartmann, 1974; Keselman & Leventhal, 1974) hat sich wegen unkorrekter Berücksichtigung der Abhängigkeit der Residuen z.T. heftiger Kritik ausgesetzt gesehen (Thoresen & Elashoff, 1974; Levin, Marascuilo & Hubert, 1978; Gottman & Glass, 1978; Bortz, 1977).

### *3.1.1 Verteilungsfreie Prüfmethode: Randomisierungs- bzw. Permutationstests*

Einfache Auswertungsverfahren, die speziell für Zeitreihenexperimente geeignet sind, bieten sich mit den Randomisierungs- oder Permutationstests an (Edgington, 1967, 1969a,b, 1971, 1973, 1975a,b, 1980). Dabei geht die prinzipielle Konstruktionsidee auf Fisher (1951<sup>6</sup>) zurück. Während die Prüfgrößen z.T. völlig äquivalent zu denen der klassischen statistischen Verfahren (t- oder F-Bruch aber auch einfache Mittelwerts- oder Mediandifferenzen) sein können, weichen Nullhypothesenformulierung und Inferenzmodell doch wesentlich vom klassischen Verfahren ab.

Beschränken wir uns hier auch wieder auf das intensive Design. Die Nullhypothese: „Kein Unterschied zwischen A und B Phasen“ wird geprüft im Vergleich mit einer bedingten Wahrscheinlichkeitsverteilung, die sich durch alle theoretisch möglichen Aufteilungen dieser gerade erhobenen Daten auf die Treatments A und B ergibt. So kann z.B. ein Experiment nach der Folge ABBAABA ablaufen. Sind wir an der Effektivität der B-Phasen interessiert, läßt sich d-unterschied zwischen A und B z.B. durch die Mittelwertsdifferenz  $\bar{Y}_B - \bar{Y}_A$  quantifizieren. Zur Signifikanzprüfung stellen wir fest, ob eine gleich große oder größere Mittelwertsdifferenz unter der Nullhypothese unwahrscheinlich ( $p \leq 0.05$ ) ist. Das hängt davon ab, inwieweit andere Aufteilungen der Daten in A und B Phasen eine mindestens ebenso große Mittelwertsdifferenz erbringen. Inferenz beschäftigt sich nicht mehr wie bei klassischen Experimenten mit der Generalisation über Personen und Situationen sondern mit der Generalisation über andere A vs. B Zuweisungen (Baseline vs. Treatmentzuweisungen). Die Nullhypothese wird beibehalten, wenn einfache Umbenennungen der experimentellen Phasen gleiche oder größere Unterschiede (hier: Mittelwertsdifferenzen) häufig auftreten lassen.

Lehmann (1975) bezeichnete dieses Inferenzmodell „Randomisierungsmodell“, das sich logisch vom üblichen „Stichproben/Populationsmodell“ abhebt. Es werden keine Populationsannahmen gemacht, da die beobachteten Daten die Rolle der Population übernehmen.

Wir wollen die Logik eines randomisierten Mittelwertsvergleichs an einem numerischen Beispiel von Kazdin (1976) demonstrieren. Statt Mittelwerte können auch andere Maße der zentralen Tendenz (wie Mediane, Proportionen etc.) Verwendung finden. Es werden folgende Daten beobachtet:

1.	2.	3.	4.	5.	6.	7.	8.	Tag
A <sub>1</sub>	B <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	B <sub>2</sub>	A <sub>4</sub>	B <sub>3</sub>	B <sub>4</sub>	Baseline A / Treatment B
20	50	15	10	60	25	65	70	abhängige Var. Y <sub>t</sub>
Y <sub>A</sub> = 17.5				Y <sub>B</sub> = 61.25				Mittelwerte
Y <sub>B</sub> − Y <sub>A</sub> = 43.75								Differenz

Zur Prüfung der einseitigen Fragestellung nach Überlegenheit von B muß untersucht werden, wie oft andere Aufteilungen der Daten in zwei Gruppen größere Mittelwertsdifferenzen liefern. Es lassen sich  $(n_A + n_B)! / (n_A! n_B!) = 8! / (4! 4!) = 70$  Aufteilungen (Permutationen) der 8 Datenwerte in jeweils zwei 4er Gruppen A' und B' finden. Die vier Aufteilungen mit den größten Mittelwertsdifferenzen sind:

	1.	2.	3.	4.	5.	6.	7.	8.	Tag	$\overline{Y}_{A'}$	$\overline{Y}_{B'}$	Differenz
1. Aufteilung	$A'_1$	$B'_1$	$A'_2$	$A'_3$	$B'_2$	$A'_4$	$B'_3$	$B'_4$	17.50	61.25	43.75	
2. Aufteilung	$A'_1$	$A'_2$	$A'_3$	$A'_4$	$B'_1$	$B'_2$	$B'_3$	$B'_4$	23.75	55.00	31.25	
3. Aufteilung	$B'_1$	$A'_1$	$A'_2$	$A'_3$	$B'_2$	$A'_4$	$B'_3$	$B'_4$	25.00	53.75	28.75	
4. Aufteilung	$A'_1$	$B'_1$	$A'_2$	$A'_3$	$A'_4$	$B'_2$	$B'_3$	$B'_4$	26.25	52.50	26.25	

Da die Wahrscheinlichkeit, durch andere Aufteilungen der Daten gleiche oder größere Mittelwertsdifferenz als die beobachtete zu erhalten, kleiner als  $p=0.05$  ist (nämlich  $1/70=0.014$ ) wird die Nullhypothese verworfen.

Die Reduktion der Daten zu *Phasenmittelwerten* (oder anderen Summationsmaßen) bringt, wie Levin et al. (1978) gezeigt haben, beträchtliche Vorteile bei der Analyse autokorrelierter Daten. Die Mittelwerte weisen eine stark reduzierte Autokorrelation gegenüber der Originalzeitreihe auf. Besitzt der stochastische Prozeß der Rohdaten z.B. folgende Autokorrelationsfunktion  $q(Y_{t_i}, Y_{t_j}) = \rho^{|i-j|} \geq 0.0$ , ist die Autokorrelation erster Ordnung für Mittelwerte, die aus 6 Einzeldaten pro Phase A oder B berechnet wurden, gleich .094, wenn die Autokorrelation der Rohdaten gleich .40 war. Ist die Autokorrelation erster Ordnung der Variablen gleich .70, ist die Autokorrelation der Mittelwerte, die aus 9 aufeinander folgenden Einzelwerten berechnet wurden, nur noch .199.

Wir wollen an einem Beispiel die Auswertung des häufig verwendeten  $A_1B_1A_2B_2$ -Designs mit Mittelwerten demonstrieren. Obwohl wir hier einen Mittelwertsvergleich anstellen, sind wir nicht darauf beschränkt. Andere Zentralitätsmaße (Mediane, Proportionen etc.) führen zur gleichen Auswertungs- und Testprozedur. Im Experiment wurden folgende Daten beobachtet:

1.	2.	3.	4. Phase	
$A_1$	$B_1$	$A_2$	$B_2$	
2.0	7.0	3.0	8.0	Mittelwerte $\bar{Y}_t$
$(7.5 - 2.5) = (\bar{Y}_B - \bar{Y}_A) = 5.0$				Differenz

Zur Überprüfung der einseitigen Fragestellung, ob  $\bar{Y}_B$  signifikant größer ist als  $\bar{Y}_A$ , muß untersucht werden, wie oft andere Aufteilungen der Daten in zwei Gruppen mindestens ebenso große Mittelwertsdifferenzen liefern. Es lassen sich  $(n_A + n_B)! / (n_A! n_B!) = 4! / (2! 2!) = 6$  Aufteilungen (Permutationen) finden:



Aufteilung	1.	2.	3.	4.	Phase	$\overline{Y}_{A'}$	$\overline{Y}_{B'}$	$\overline{Y}_{B'} - \overline{Y}_{A'}$
1	$A'_1$	$A'_2$	$B'_1$	$B'_2$		4.5	5.5	1.0
2	$A'_1$	$B'_1$	$A'_2$	$B'_2$		2.5	7.5	5.0
3	$A'_1$	$B'_1$	$B'_2$	$A'_2$		5.0	5.0	0.0
4	$B'_1$	$A'_1$	$A'_2$	$B'_2$		5.0	5.0	0.0
5	$B'_1$	$A'_1$	$B'_2$	$A'_2$		7.5	2.5	-5.0
6	$B'_1$	$B'_2$	$A'_1$	$A'_2$		5.5	4.5	-1.0

Da die Wahrscheinlichkeit, durch eine andere Aufteilung eine mindestens ebenso große Mittelwertsdifferenz wie die beobachtete zu finden, größer als .05 ist (nämlich exakt 1/6), wird die Nullhypothese beibehalten. Da dieses  $p=1/6=.16$  nach den üblichen Standards zu groß ist, kann das  $A_1B_1A_2B_2$ -Design nicht über Mittelwertvergleiche innerhalb des Randomisierungsinferenzmodell ausgewertet werden! Fügt man eine weitere Baselinephase an ( $A_1B_1A_2B_2A_3$ -Design), lassen sich zwar  $5!/(3!2!) = 10$  Aufteilungen der 5 Mittelwerte finden. Das kleinste  $p$  ist aber im günstigsten Fall bei einseitiger Fragestellung ebenfalls mit  $p=1/10=.10$  zu hoch.

### *Ordinale Gewichtungsschemata bei Randomisierungstests*

Um ein noch kleineres  $p$  zu erhalten, müssen gerichtete Hypothesen geprüft werden. Eine gerichtete Alternativhypothese ist z.B.:

$$\mu_{A_1} < \mu_{A_2} < \mu_{B_1} < \mu_{B_2} \text{ oder: } \mu_{A_1} < \mu_{A_2} < \{\mu_{B_1}, \mu_{B_2}\}$$

Aus dieser Hypothese leitet sich das ordinale Gewichtungsschema ab: die  $\bar{Y}$  werden entsprechend der prognostizierten Größe mit ihren Rängen gewichtet:

$$g_{A_1} = 1, g_{A_2} = 2, g_{B_1} = 3, g_{B_2} = 4$$

Mit diesen Gewichten werden die empirisch erhaltenen Maße (Mittelwerte, Mediane, Proportionen) multipliziert und aufsummiert:

$$\bar{Y}_{A_1} = 2.0, \bar{Y}_{A_2} = 3.0, \bar{Y}_{B_1} = 7.0, \bar{Y}_{B_2} = 8.0$$

$$\Gamma_{\text{emp}} = \sum_{t=1}^4 g_t \bar{Y}_t = 1 \times 2.0 + 2 \times 3.0 + 3 \times 7.0 + 4 \times 8.0 = 61.0$$

Es bleibt zu prüfen, wie groß die Auftretenswahrscheinlichkeit einer mindestens ebenso großen Summe  $\Gamma$  ist, wenn man die Zuordnung (Benennung) der Maße zu den Phasen  $A_1 - B_2$  permutiert. Die Verteilung der gewichteten Summen  $\Gamma_k$  findet sich in Tabelle 3.1.

Tabelle 3.1: Verteilung gewichteter Summen  $\Gamma_k$  unter der Nullhypothese in einem ABAB-Design

Permutation	1.	2.	3.	4.	Experimentelle Phase	$\Gamma$
K	2.0	7.0	3.0	8.0	Mittelwerte	
1	A <sub>1</sub> '	A <sub>2</sub> '	B <sub>1</sub> '	B <sub>2</sub> '		57
2	A <sub>1</sub> '	A <sub>2</sub> '	B <sub>2</sub> '	B <sub>1</sub> '		52
3	A <sub>1</sub> '	B <sub>1</sub> '	A <sub>2</sub> '	B <sub>2</sub> '		61
4	A <sub>1</sub> '	B <sub>1</sub> '	B <sub>2</sub> '	A <sub>2</sub> '		51
5	A <sub>1</sub> '	B <sub>2</sub> '	A <sub>2</sub> '	B <sub>1</sub> '		60
6	A <sub>1</sub> '	B <sub>2</sub> '	B <sub>1</sub> '	A <sub>2</sub> '		55
7	A <sub>2</sub> '	A <sub>1</sub> '	B <sub>1</sub> '	B <sub>2</sub> '		52
8	A <sub>2</sub> '	A <sub>1</sub> '	B <sub>2</sub> '	B <sub>1</sub> '		47
9	A <sub>2</sub> '	B <sub>1</sub> '	A <sub>1</sub> '	B <sub>2</sub> '		60
10	A <sub>2</sub> '	B <sub>1</sub> '	B <sub>2</sub> '	A <sub>1</sub> '		45
11	A <sub>2</sub> '	B <sub>2</sub> '	A <sub>1</sub> '	B <sub>1</sub> '		59
12	A <sub>2</sub> '	B <sub>2</sub> '	B <sub>1</sub> '	A <sub>1</sub> '		49
13	B <sub>1</sub> '	A <sub>1</sub> '	A <sub>2</sub> '	B <sub>2</sub> '		51
14	B <sub>1</sub> '	A <sub>1</sub> '	B <sub>2</sub> '	A <sub>2</sub> '		41
15	B <sub>1</sub> '	A <sub>2</sub> '	A <sub>1</sub> '	B <sub>2</sub> '		55
16	B <sub>1</sub> '	A <sub>2</sub> '	B <sub>2</sub> '	A <sub>1</sub> '		40
17	B <sub>1</sub> '	B <sub>2</sub> '	A <sub>1</sub> '	A <sub>2</sub> '		53
18	B <sub>1</sub> '	B <sub>2</sub> '	A <sub>2</sub> '	A <sub>1</sub> '		48
19	B <sub>2</sub> '	A <sub>1</sub> '	A <sub>2</sub> '	B <sub>1</sub> '		45
20	B <sub>2</sub> '	A <sub>1</sub> '	B <sub>1</sub> '	A <sub>2</sub> '		40
21	B <sub>2</sub> '	A <sub>2</sub> '	A <sub>1</sub> '	B <sub>1</sub> '		49
22	B <sub>2</sub> '	A <sub>2</sub> '	B <sub>1</sub> '	A <sub>1</sub> '		39
23	B <sub>2</sub> '	B <sub>1</sub> '	A <sub>1</sub> '	A <sub>2</sub> '		48
24	B <sub>2</sub> '	B <sub>1</sub> '	A <sub>2</sub> '	A <sub>1</sub> '		43

Da die beobachtete Reihe von Mittelwerten in einer ordinal gewichteten Summe  $\Gamma$  von 61 resultiert und die Wahrscheinlichkeit, unter der Nullhypothese eine mindest ebenso große Summe zu erhalten, gering ist, (s.a. rechte Spalte in Tab. 3.1) wird  $H_0$  zugunsten von  $H_1$  verworfen: Die Phasen unterscheiden sich signifikant.

#### *Nichtordinale Gewichtungsschemata und Kontraste*

Statt eines ordinalen Gewichtsschema ist auch ein Intervall- oder Verhältnisschema denkbar, je nachdem wie präzise die Vorhersagen über die empirischen

Maße formuliert werden. Differenzen (Verhältnisse) zwischen den Gewichten entsprechen den erwarteten oder prognostizierten Differenzen (Verhältnissen) in den Summationsmaßen.

Andere Gewichtungsschemata werden als lineare Kontraste verwendet. Ein Beispiel für eine verbal formulierte Hypothese findet sich bei Elashoff & Thorensen (1978, S. 308). Besteht das Experiment z.B. aus fünf Phasen  $A_1B-C_1A_2C_2$  und prognostiziert man (Alternativhypothese  $H_1$ ), daß einerseits die Maße folgende Rangreihe aufweisen:

$$\{\mu_{A_1}, \mu_{A_2}\} < \mu_B < \{\mu_{C_1}, \mu_{C_2}\}$$

und andererseits der Unterschied der A Phasen zur B Phase genauso groß ist wie der Unterschied der C-Phasen zur B-Phase, muß folgende Gleichung gelten

$$((\mu_{A_1} - \mu_B) + (\mu_{A_2} - \mu_B)) = ((\mu_{C_1} - \mu_B) + (\mu_{C_2} - \mu_B))$$

bzw.

$$1 \mu_{A_1} + 1 \mu_{A_2} - 1 \mu_{C_1} - 1 \mu_{C_2} = 0$$

Das Gewichtungsschema lautet dann:

$$g_{A_1} = 1, g_B = 0, g_{C_1} = -1, g_{A_2} = 1, g_{C_2} = -1$$

Die Signifikanz wird dann nach dem üblichen Schema beurteilt: a) Berechnung des  $\Gamma_{\text{emp}}$ , der Verteilung der  $\Gamma_k$  unter der Nullhypothese über Permutationen, b) Vergleich des empirisch gewonnenen  $\Gamma_{\text{emp}}$  mit der über Permutationen abgeleiteten Prüfverteilung.

### Approximierte Prüfverteilung beim Randomisierungstest

Liegen mehr als vier Phasen vor, wird die Ableitung der exakten Verteilung von  $\Gamma$  aufwendig, da die Zahl der zu beachtenden Permutationen stark ansteigt. Bei mindestens vier A-Phasen ( $A_1A_2A_3A_4$ ) und vier B-Phasen ( $B_1B_2B_3B_4$ ) ist die durch ein Gewichtungsschema bestimmte Summe  $\Gamma_k = \sum g_{ik} \bar{Y}_t$  annähernd normalverteilt (Levin, Marascuilo & Hubert, 1978). Die Signifikanz von  $\Gamma_{\text{emp}}$  läßt sich an Hand des folgenden z-Wertes prüfen:

$$(3.1) \quad Z_{\text{emp}} = \frac{(\Gamma_{\text{emp}} - E(\Gamma))}{\sqrt{\text{Var}(\Gamma)}}$$

$$\text{mit } E(\Gamma) \approx \frac{1}{n} \left( \sum_{t=1}^n g_t \right) \left( \sum_{t=1}^n \bar{Y}_t \right)$$

$$\text{und Var } (\Gamma) \approx \frac{1}{n-1} \left[ \sum_{t=1}^n (g_t - \bar{g})^2 \right] \left[ \sum (Y_t - \bar{Y})^2 \right]$$

Als Beispiel soll ein Experiment von acht Phasen dienen ( $A_1, A_2, A_3, A_4, B_1, B_2, B_3, B_4$ -Design). Es wurden die Mittelwerte

$$\bar{Y}_{A_1} = .1, \bar{Y}_{A_2} = .3, \bar{Y}_{A_3} = .4, \bar{Y}_{A_4} = .8, \bar{Y}_{B_1} = .2, \bar{Y}_{B_2} = .5, \bar{Y}_{B_3} = .6, \bar{Y}_{B_4} = .7$$

beobachtet. Die vor dem Experiment aufgestellte Alternativhypothese  $H_1$

$$A_1 < A_2 < A_3 < A_4 < B_1 < B_2 < B_3 < B_4$$

führt zum ordinalen Gewichtungsschema:

$$g_{A_1} = 1, g_{A_2} = 2, g_{A_3} = 3, g_{A_4} = 4, g_{B_1} = 5, g_{B_2} = 6, g_{B_3} = 7, g_{B_4} = 8$$

und zur gewichteten Summe  $\Gamma = \sum_{t=1}^8 g_t \bar{Y}_t = 18.9$ . Der z-Wert ist

$$z = \left[ 18.9 - \frac{36 \times 3.6}{8} \right] / \sqrt{\frac{42 \times 0.42}{7}} = 1.702$$

Bei einseitiger Fragestellung auf dem 5% Niveau ist dieses Ergebnis nicht signifikant.

### 3.1.2 Verteilungsgebundene Prüfverfahren: Lineares Modell

Bei Verwendung linearer Auswertungsprozeduren (t-Test, Varianzanalyse etc.) zur Prüfung von Veränderungen bei Längsschnittdaten ist wegen der möglichen seriellen Abhängigkeit der Daten Vorsicht geboten, weil die kritischen t- oder F-Brüche „zu groß“ sein können.

Beschränken wir uns hier wieder auf den Mittelwertsvergleich in einem Vorher-Nachher-Design (AB Design). Wären die Daten unabhängig voneinander erhoben (z.B. zwei unabhängige Stichproben für die A- und die B-Phase) könnte man den normalen t-Test mit Hilfe des allgemeinen linearen Modells formulieren. Dabei haben wir von den verschiedenen Möglichkeiten zur Gestaltung der Designmatrix  $X$  eine Form gewählt, die für die Prüfung eines Niveaustiegs besonders sinnvoll erscheint. Wir wählen für das volle Modell:

(3.1.2a)

$Y_{A1}$  $\vdots$  $Y_{Ai}$  $\vdots$  $Y_{AN_A}$

$Y_{B1}$  $\vdots$  $Y_{Bj}$  $\vdots$  $Y_{BN_B}$

=

1

0

1

1

$\beta_1$  $\beta_2$

+  $\varepsilon_v$

$Y_{Ai}$  = Wert der abhängigen Variablen für die Person i in der Gruppe A

$Y_{Bj}$  = Wert der abhängigen Variablen für die Person j in der Gruppe B

oder kurz

(3.1.2b)

$y = X_v \beta_v + \varepsilon_v$

mit Fehlerquadratsumme

(3.1.3)

$F_v^2 = \hat{\varepsilon}_v' \hat{\varepsilon}_v$

Weisen die Gruppen hinsichtlich der Mittelwerte keine Unterschiede auf, darf der Niveauanstiegparameter  $\beta_2$  nicht signifikant sein.

Statt  $\beta_2$  direkt auf Signifikanz zu prüfen, können wir aus dem unter der Alternativhypothese formulierten vollen Modell durch Einführung der Nullhypothese ein reduziertes Modell mit entsprechender Fehlerquadratsumme Ff herleiten. Die Nullhypothese lautet  $\beta_2=0$  (bzw.  $\mu_A=\mu_B$ ). Das reduzierte Modell nimmt folgende Gestalt an

(3.1.4a)

$y_A$

$y_B$

=

1

1

$\beta_r$

+  $\varepsilon_r$

oder kurz

(3.1.4b)

$y = X_r \beta_r + \varepsilon_r$

mit Fehlerquadratsumme

$$(3.1.5) \quad F_r^2 = \hat{\varepsilon}'_r \hat{\varepsilon}_r$$

Die Nullhypothese läßt sich über einen F-Bruch prüfen, der in diesem 2-Gruppenfall zu den gleichen Schlüssen führt wie der übliche t-Test:

$$F_{df1, df2} = \frac{(F_r^2 - F_v^2)/df_1}{(F_v^2 - 0.0)/df_2} \quad \text{mit} \quad \begin{array}{l} df_1 = p_v - p_r \\ df_2 = N - p_v \end{array}$$

(3.1.6)  $p_v$  = Zahl der linear unabhängigen Prädiktoren im vollen Modell (hier: 2)

$p_r$  = Zahl der linear unabhängigen Prädiktoren im reduzierten Modell (hier: 1)

$N = (N_1 + N_2)$  = Größe der Gesamtstichprobe

Die Angemessenheit dieses Vorgehens hängt von der Gültigkeit der drei Annahmen (3.1.7) ab (s. Searle, 1971; Timm, 1975; Wottawa, 1974):

$$(3.1.7a) \quad y = X\beta + \varepsilon \quad (\text{Gültigkeit des Linearmodells})$$

$$(3.1.7b) \quad E(\varepsilon) = 0$$

$$(3.1.7c) \quad E(\varepsilon\varepsilon') = \sigma_\varepsilon^2 I \quad (\text{Unabhängigkeit der Residuen})$$

mit  $I$  = Einheitsmatrix

Insbesondere (3.1.7c) ist eine für Längsschnittdaten kritische Annahme. Sie wird bei Querschnittdaten nicht überprüft, da man sich auf die Wirksamkeit der Randomisierung verläßt. Für den Fall abhängiger Messungen ist sie jedoch in der Regel verletzt.

Wollen wir den Niveauanstieg in der B-Phase nicht mehr wie in (3.1.2) und (3.1.4) mit zwei unabhängigen Stichproben sondern mit einer Stichprobe, die zweimal getestet wurde, würden sich in die Modelle (3.1.2) und (3.1.4) Paare voneinander abhängiger Residuen einschleichen. Hat nämlich die Person  $i$  im Durchgang A einen viel höheren (niederen) Wert  $Y_{iA}$  als der Durchschnitt der Gruppe (bzw. als der Gesamtdurchschnitt), sind die zu dieser Person gehörenden Residuen  $\varepsilon_{Ai}$  und  $\varepsilon_{Bi}$  einander ähnlich (d.h. sie korrelieren), weil wir erwarten, daß der zur Person gehörende 2. Meßwert  $Y_{Bi}$  auch über (bzw. unter) dem Durchschnitt bleibt.

Die diagonale Kovarianzmatrix der Residuen (3.1.7c) muß noch um mindestens einen Korrelationsparameter  $\varrho$  erweitert werden zu (3.1.8) (S. 296).

Ignoriert man die Abhängigkeit der Residuen, werden bei positiver Korreliertheit die F-Brüche und damit der  $\alpha$ -Fehler zu groß: die Nullhypothese wird zu oft verworfen. Man kann auf drei Arten dieser Verfälschung entgegen: (a) Korrektur der Freiheitsgrade beim F-Bruch (Box, 1954; Geisser & Greenhouse, 1958; McCall & Applebaum, 1973), (b) Aufspaltung der korrelierten Resi-

$$(3.1.8) \quad E(\varepsilon \varepsilon') = \sigma_\varepsilon^2 \Omega = \sigma_\varepsilon^2 \cdot \begin{array}{cc|c} \begin{array}{cc|c} 10 & \dots & 0 \\ 01 & \dots & 0 \\ \hline 0 & \dots & 1 \end{array} & \begin{array}{cc|c} \rho 0 & \dots & 0 \\ 0 \rho & \dots & 0 \\ \hline 0 & \dots & \rho \end{array} & \begin{array}{c} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ N \end{array} \\ \hline \begin{array}{cc|c} \rho 0 & \dots & 0 \\ 0 \rho & \dots & 0 \\ \hline 0 & \dots & \rho \end{array} & \begin{array}{cc|c} 10 & \dots & 0 \\ 01 & \dots & 0 \\ \hline 0 & \dots & 1 \end{array} & \begin{array}{c} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ N \end{array} \end{array} \neq \sigma_\varepsilon^2 I$$

$\left. \begin{array}{c} \text{A} \\ \text{B} \end{array} \right\}$

duen in Personeneffekte und unkorrelierte neue Fehler durch Erweiterung der Designmatrix  $X_v$  und  $X_r$  um Personenvektoren (McNeil, Kelly & McNeil, 1975; Pedhazur, 1977) und schließlich (c) multivariate Auswertung (Bock, 1975, 1979; Finn, 1968; Timm, 1975).

Ist man aber aus den verschiedensten Gründen nicht daran interessiert, die Effektivität der Intervention B an mehreren Personen zu prüfen, wendet man sich wieder dem intensiven Design (bzw. dem  $N=1$  Experiment) zu. Wird an einer Person eine Reihe von Meßwerten (= Zeitreihe) beobachtet, kann die Abhängigkeitsstruktur der Residuen wesentlich komplizierter aussehen als z.B. in (3.1.8)\*. So ist einer der allgemeinsten Fälle der theoretischen Abhängigkeitsstruktur (bei homogener Fehlervarianz  $\sigma_\varepsilon^2 = \sigma_\varepsilon^2$ ):

$$(3.1.9) \quad E(\varepsilon \varepsilon') = \sigma_\varepsilon^2 \Omega = \sigma_\varepsilon^2 \cdot \begin{array}{cc|c} \begin{array}{ccc|c} 1 & \rho_1 & \rho_2 & \dots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & & \\ \rho_2 & \rho_1 & 1 & & \\ \rho_3 & \rho_2 & & \ddots & \\ & & 1 & & \\ & & \rho_1 & 1 & \\ & & \rho_2 & \rho_1 & 1 \\ & & & & \ddots & \\ \rho_{T-1} & & & & & \rho_3 & \rho_2 & \rho_1 & 1 \end{array} & \begin{array}{c} 1 \\ 2 \\ \vdots \\ t \\ \vdots \\ T \end{array} \end{array}$$

$\left. \begin{array}{c} \text{A} \\ \text{B} \end{array} \right\}$

Beachtet man nicht, daß eventuell  $\sigma_\varepsilon^2 \Omega \neq \sigma_\varepsilon^2 I$  ist, können auch im Zeitreihenexperiment die statistischen Schlüsse erheblich verfälscht werden. Die aus den Daten berechneten F- oder t-Werte sind „zu klein“ oder „zu groß“ (Scheffe, 1959; Gastwirth & Rubin, 1971; Glass, Peckham & Sanders, 1972; Hibbs, 1974; Gottman & Glass, 1978; Revenstorf & Keeser, 1979).

\* Ein autoregressiver Prozeß 1. Ordnung in den Residuen kann mit dem Durbin-Watson-Test geprüft werden (Makridakis & Wheelwright, 1978a,b)

Als Ausweg bietet sich an, die Modelle (3.1.2) und (3.1.4) mit einer zunächst noch unbekannten Matrix  $A$  so zu transformieren, daß die neuen Residuen  $\varepsilon^+$  wieder unabhängig sind (Hibbs, 1974):

$$(3.1.10a) \quad Ay = AX\beta + A\varepsilon$$

$$(3.1.10b) \quad y^+ = X^+ \beta + \varepsilon^+ \text{ transformiertes Linearmodell}$$

$$(3.1.10c) \quad \Omega^+ = A \Omega A' = I \quad \text{transformierte Korrelationsmatrix der Residuen}$$

$$(3.1.10d) \quad E(\varepsilon^+ \varepsilon^{+'}) = E(A\varepsilon \varepsilon' A') = A E(\varepsilon \varepsilon') A' = \sigma_\varepsilon^2 A \Omega A' = \sigma_\varepsilon^2 I$$

Der Schätzer

$$(3.1.11) \quad \hat{\beta} = (X^{+'} X^+)^{-1} X^{+'} y^+$$

wird auch GLS-Schätzer (generalized least squares) genannt. Diese Regressionsmethode geht auf Aitken (1935) zurück.

Problematisch an der GLS-Methode ist die Schätzung der Matrix  $\Omega$ . Erst wenn man sie geschätzt hätte, ließe sich die Transformationsmatrix nach (3.1.10c) bestimmen. Da die Schätzung der  $Q_t$  bei einer endlichen Zeitreihe mit der Länge  $T$  immer ungenauer wird, je höher die Ordnung der Autokorrelation der Residuen ist, nimmt man für die Residuen einfache ARIMA-Prozesse an. Hierfür gibt Hibbs (1974) verschiedene Transformationsmatrizen  $A$  an.

Praktikabilität erreichte die Interventionsanalyse erst nach einem Artikel von Box & Tiao (1965), der von Glass, Willson & Gottman (1975) speziell für die praktische Einzelfalldiagnostik ausgearbeitet wurde. Da die stochastische Struktur der Residuen nicht bekannt ist, untersuchen Box & Tiao sowie Gottman et al. zuerst die stochastische Struktur der Rohwerte. In einem iterativen Prozeß werden dann ähnlich wie in (3.1.10a) eine Transformation des Datenvektors  $y$  und der Designmatrix  $X$  gesucht, so daß die Residuen den Annahmen des allgemeinen linearen Modells (3.1.7c) entsprechen. Wir wollen an drei Beispielen diese Art der Interventionstestung darstellen.

Folgen die Rohdaten einem moving-average-Prozeß 1. Ordnung (2.13) (ARIMA(0,0,1)) und unterscheiden sich A- und B-Phase nur in ihrem Niveau, ist das Zeitreihenmodell *vor* der Intervention:

$$(3.1.12) \quad y_t = L - \theta_1 a_{t-1} + a_t = f(L, \theta_1, a_{t-1}, a_t) \quad a_t \sim N(0, \sigma^2)$$

eine Funktion eines Levelparameters  $L$ , des moving-average-Parameters  $\theta_1$  und der gegenwärtigen und vergangenen „Fehler“ oder Schocks  $a_t, a_{t-1}$ .

Der von Glass, Willson & Gottman verwendete Levelparameter  $L$  trägt zweierlei Bedeutung. Bei der Formulierung der ARIMA(p,d,q)-Prozesse spielt er die Rolle von  $\mu_y = E(Y_t)$  bzw.  $\mu_W = E(W_t)$  (s.a. Kap. 3). Bei der Transforma-



tion (3.1.10) nehmen die Autoren vereinfachend an, daß alle  $a_t = 0$  für  $t < 1$  ist, so daß  $Y_t = L$  für  $t < 1$  ist.

$L$  wäre dann das Niveau der Zeitreihe zum Zeitpunkt  $t=0$ . Die Annahme  $a_t = 0$  für  $t < 1$  wirkt sich natürlich um so schwächer aus, je länger die Zeitreihe ist.

Das Modell nach dem Interventionsbeginn:

$$(3.1.13) \quad Y_t = L + \delta - \theta_1 a_{t-1} + a_t = f(L, \delta, \theta_1, a_t, a_{t-1})$$

unterscheidet sich von (3.1.12) nur durch den Parameter  $\delta$ , der den Niveauunterschied der Zeitreihen in der A- und B-Phase widerspiegelt. Unter der Nullhypothese ist  $\delta$  gleich Null. Für die Auswertung nach dem allgemeinen linearen Modell (3.1.2), (3.1.4) und (3.1.6) stört die Abhängigkeit des Meßwerts  $Y_t$  vom vergangenen Fehler  $a_{t-1}$ . Wir müssen daher neue transformierte Werte  $Y_t^+ = f(Y_t)$  finden, in denen  $a_{t-1}$  nicht mehr enthalten ist. Die neuen Werte  $Y_t^+$  sollen nur noch von  $L, \delta, \theta_1, a_t$  abhängen:

$$(3.1.14) \quad Y_t^+ = f(Y_t) = f(L, \delta, \theta_1, a_t)$$

Wir suchen also ähnlich zu (3.1.10) eine Transformationsmatrix  $A$ . Nimmt man an, daß die  $a_t$  für Zeiten vor der ersten Beobachtung  $Y_1$  Null sind:  $a_t = 0$  für  $t < 1$ , lautet diese Matrix (s. Revenstorff & Keeser, 1978).

$$(3.1.15) \quad A = \begin{bmatrix} 1 & 0 & 0 & 0 \dots 0 & 1 \\ \theta_1 & 1 & 0 & 0 \dots 0 & 2 \\ \theta_1^2 & \theta_1 & 1 & 0 \dots 0 & \vdots \\ \theta_1^3 & \theta_1^2 & \theta_1 & 1 & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_1^{T-1} & \theta_1^{T-2} & \theta_1^{T-3} & \dots & 1 & T \end{bmatrix}$$

Zur Hypothesentestung werden wieder zwei Regressionen gerechnet. Für das volle Modell mit dem Ansatz:

$$(3.1.16a) \quad \begin{array}{l} \text{Phase A} \\ Y_1^+ \\ Y_2^+ \\ Y_3^+ \\ \vdots \\ Y_n^+ \end{array} = \begin{bmatrix} 1 & 0 \\ (1 + \theta_1) & 0 \\ (1 + \theta_1 + \theta_1^2) & 0 \\ \vdots & \vdots \\ (1 + \theta_1 + \dots + \theta_1^{n-1}) & 0 \end{bmatrix} \begin{bmatrix} L \\ \delta \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

$$\begin{array}{l} \text{Phase B} \\ Y_{n+1}^+ \\ Y_{n+2}^+ \\ \vdots \\ Y_T^+ \end{array} = \begin{bmatrix} (1 + \theta_1 + \dots + \theta_1^{n-1} + \theta_1^n) & 1 \\ (1 + \theta_1 + \dots + \theta_1^n + \theta_1^{n+1}) & (1 + \theta_1) \\ \vdots & \vdots \\ \vdots & \vdots \\ (1 + \theta_1 + \dots + \theta_1^{T-1}) & (1 + \theta_1 + \dots + \theta_1^{T-n-1}) \end{bmatrix} \begin{bmatrix} L \\ \delta \end{bmatrix} + \begin{bmatrix} a_{n+1} \\ \vdots \\ a_T \end{bmatrix}$$

oder kurz

$$(3.1.16b) \quad y^+ = (X_1^+ | X_2^+) \begin{pmatrix} L \\ \delta \end{pmatrix} + a_{\text{voll}}$$

mit Fehlerquadratsumme

$$(3.1.17) \quad F_v^2(L, \delta, \theta_1) = \hat{a}_{\text{voll}}' \hat{a}_{\text{voll}}$$

und für das reduzierte Modell:

$$(3.1.18) \quad y^+ = (X_1^+) \cdot L + a_{\text{red}}$$

mit Fehlerquadratsumme:

$$(3.1.19) \quad \hat{a}_r' \hat{a}_r = F_r^2(L, \theta_1)$$

Die Nullhypothese  $\delta = 0$  wird mit dem üblichen F-Bruch getestet:

$$(3.1.20) \quad F_{df1, df2} = \frac{(F_r^2 - F_v^2)/df_1}{(F_v^2 - 0)/df_2} = \frac{(F_r^2 - F_v^2)/(2-1)}{F_v^2/(T-2)} = t_{T-2}^2$$

Glass, Willson & Gottman (1975, S. 125ff.) bringen als Beispiel für ein Zeitreihenexperiment mit AB-Phasen eine Interventionsevaluation, bei dem die Zeitreihe des störenden Schülerverhaltens einem ARIMA(0,0,1)-Prozeß folgte (s. Figur 3.1.3). Nach dem 21. Tag wurde versucht, das Schülerverhalten zu ändern (Hall et al., 1971). Die geschätzten Parameterwerte betrugen  $\hat{L} = 19.24$ ,  $\hat{\delta} = -14.29$ ,  $\theta_1 = -.34$  und der  $t_{df=38} = [F_{1,38}]^{1/2}$  Wert belief sich auf -16.39, was hoch signifikant war.

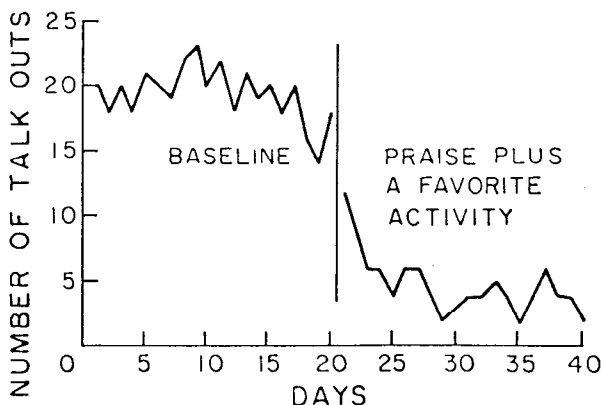


Fig. 3.1.3: Anzahl von Unterrichtsstörungen durch Schwatzen der Schüler vor und nach Interventionsbeginn

Folgen die Rohdaten einem autoregressiven Prozeß 1. Ordnung (s. 2.8d, 2.9a) ARIMA(1,0,0) ist das Zeitreihenmodell *vor* der Intervention:

$$(3.1.21a) \quad Y_t - L = \varphi_1(Y_{t-1} - L) + a_t \quad -1 < \varphi_1 < 1$$

$a_t \sim N(0, \sigma^2)$  weißes Rauschen  
 $t = 1, \dots, N_1 = n$

oder

$$(3.1.21b) \quad Y_t = \varphi_1 Y_{t-1} + L(1 - \varphi_1) + a_t$$

und das Modell nach dem Interventionsbeginn

$$(3.1.22a) \quad Y_t - (L + \delta) = \varphi_1(Y_{t-1} - (L + \delta)) + a_t$$

oder

$$(3.1.22b) \quad Y_t = \varphi_1 Y_{t-1} + (L + \delta)(1 - \varphi_1) + a_t$$

Auch hier ist wieder eine neue Zeitreihe  $Y_t^+ = f(Y_t)$  gesucht, die nicht mehr von den das allgemeine lineare Modell störenden  $Y_{t-1}$  abhängt. Die gesuchte Transformationsmatrix ist (bei Annahme, daß die  $a_t = 0$  für  $t < 1$  und  $Y_0 = L$ ; s. Revenstorff & Keeser, 1978, S. 20).

$$(3.1.23) \quad A = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -\varphi_1 & 1 & 0 & \dots & 0 \\ 0 & -\varphi_1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & -\varphi_1 & 1 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ \vdots \\ \vdots \\ T \end{matrix}$$

und das volle Modell für den Test auf  $H_0: \delta = 0$

$$(3.1.24) \quad \begin{bmatrix} Y_1^+ \\ Y_2^+ \\ \vdots \\ Y_n^+ \\ Y_{n+1}^+ \\ Y_{n+2}^+ \\ \vdots \\ Y_T^+ \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ (1 - \varphi_1) & 0 \\ \vdots & \vdots \\ (1 - \varphi_1) & 0 \\ \hline (1 - \varphi_1) & 1 \\ (1 + \varphi_1) & (1 - \varphi_1) \\ \vdots & \vdots \\ (1 - \varphi_1) & (1 - \varphi_1) \end{bmatrix} \cdot \begin{bmatrix} L \\ \delta \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \\ a_{n+1} \\ a_{n+2} \\ \vdots \\ a_T \end{bmatrix}$$

Auch hierfür geben Glass, Willson & Gottman (1975) ein Beispiel (Figur 3.1.4). Die Daten stammen aus dem Projekt „Whistlestop“. Zwischen April 1969 und September 1973 wurde die Zahl der Überfälle im Hyde-Park (Chicago) registriert. Die „Intervention“ bestand darin, an die Bürger Pfeifen auszu-teilen, mit denen sie im Notfall die Polizei um Hilfe pfeifen konnten. Die Zeitreihe läßt sich mit einem ARIMA(1,0,0)-Prozeß beschreiben. Die ge-schätzten Parameter betrugen  $\hat{L} = 64.52$ ,  $\hat{\delta} = -5.41$ ,  $\hat{\phi}_1 = .50$ . Der t-Wert belief sich auf  $t_{46} = -.72$  und war somit nicht signifikant, d.h.: das Austeilen der Pfeifen hat die Zahl der Überfälle nicht signifikant reduziert.

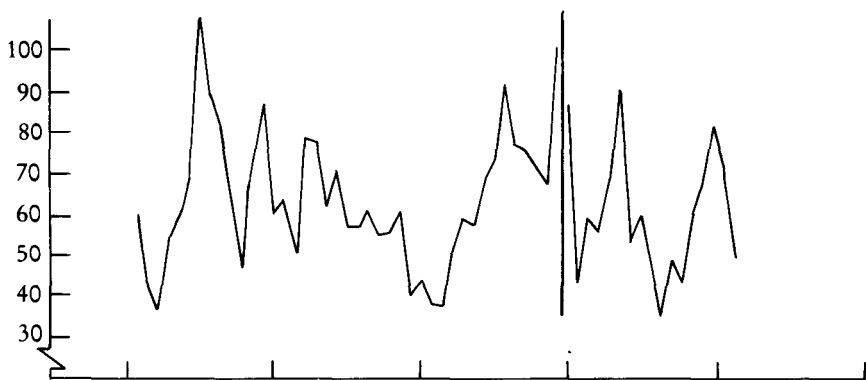


Fig. 3.1.4: Zahl der überfälle pro Monat zwischen April 1969 und September 1973 im Hyde-Park, Chicago, vor und während Projekt „Whistlestop“

Zum Schluß soll an einem ARIMA(0,1,1)-Prozeß der Rohdaten das dritte Beispiel eines Hypothesentests bezüglich des Niveaus der Zeitreihe demonstriert werden. Gottman & Glass (1978) untersuchten, ob der Wert auf einer „Irritierbarkeitsskala“ am Anfang der Menstruationsperiode einer Frau signifikant von den Werten vor und nach dem Anfang der Menstruationsperiode abweicht. Die Rohdaten sind in Figur (3.1.5) dargestellt. Würden die Residuen unabhängig von einander sein, könnte man mit einer Modifikation des allgemeinen linearen Modells (3.1.2), (3.1.4) und (3.1.6) diese Hypothese testen. Das volle Modell hätte dann die Form (3.1.25a) (S. 302)

oder kurz:

$$(3.1.25b) \quad y = (X_1 | X_2) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon_v$$

und das reduzierte Modell würde sich formulieren lassen als:

$$(3.1.25a) \quad \begin{array}{|c|} \hline Y_1 \\ Y_2 \\ \vdots \\ \hline Y_{14} \\ Y_{15} \\ \hline Y_{16} \\ \vdots \\ Y_{34} \\ \hline \end{array} = \begin{array}{|cc|} \hline 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 1 & 1 \\ 1 & 1 \\ \hline 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline \end{array} \cdot \begin{array}{|c|} \hline \beta_1 \\ \beta_2 \\ \hline \end{array} + \begin{array}{|c|} \hline \\ \hline \epsilon_v \\ \hline \end{array}$$

$$(3.1.26) \quad y = (X_1) \quad b_1 = e_t$$

Die Hypothese würde dann über den F-Bruch

$$(3.1.27) \quad F_{df_1, df_2} = \frac{(F_r^2 - F_v^2)/df_1}{(F_v^2 - 0.0)/df_2} = \frac{(F_r^2 - F_v^2)/(2-1)}{F_v^2/(T-2)} = t_{df=32}^2$$

getestet.

Jedoch erscheint die Zeitreihe nichtstationär. Das Niveau wandert erheblich während der ersten 13 Tage. Werden dagegen die Daten differenziert, d.h.  $y_t - y_{t-1}$  gebildet, scheint nach dem Autokorrelationsmuster ein ARIMA-(0,1,1)-Prozeß vorzuliegen. Nach dem ARIMA(0,1,1)-Prozeß werden die Daten  $y_t$  nach folgendem Modell vor der „Intervention“ (hier: Anfang der Menstruation) erzeugt:

$$(3.1.28a) \quad \begin{aligned} (1-B)y_t &= y_t - y_{t-1} = -\theta_1 a_{t-1} + a_t & -1 < \theta_1 < 1 \\ a_t &\sim N(0, \sigma^2) \text{ weißes Rauschen} \\ a_0 &= 0 \\ y_t &= Y_t - L \end{aligned}$$

oder:

$$(3.1.28b) \quad Y_t = L + (1-\theta_1)(a_1 + a_2 + \dots + a_{t-1}) + a_t \quad Y_0 = L$$

und während der „Intervention“ (Menstruation) nach

$$(3.1.29) \quad Y_t = L + \delta + (1-\theta_1)(a_1 + a_2 + \dots + a_{t-1}) + a_t$$

Für die Auswertung nach dem allgemeinen linearen Modell stört, daß  $Y_t$  auch von  $Y_{t-1}$  und  $a_{t-1}$  abhängt.

Das volle Modell lautet dann (Glass, Willson & Gottman, 1978, S. 316):

$$(3.1.30a) \quad \begin{array}{|c|} \hline Y_1^+ \\ Y_2^+ \\ \vdots \\ \vdots \\ \vdots \\ \hline Y_{14}^+ \\ Y_{15}^+ \\ \hline Y_{16}^+ \\ \vdots \\ Y_{34}^+ \\ \hline \end{array} = \begin{array}{|c|c|} \hline 1 & 0 \\ \theta_1 & 0 \\ \theta_1^2 & 0 \\ \vdots & \vdots \\ \theta_1^{12} & 0 \\ \hline \theta_1^{13} & 1 \\ \theta_1^{14} & \theta_1 \\ \hline \theta_1^{15} & 0 \\ \vdots & \vdots \\ \theta_1^{33} & 0 \\ \hline \end{array} \cdot \begin{array}{|c|} \hline L \\ \delta \\ \hline \end{array} + \begin{array}{|c|} \hline a_1 \\ a_2 \\ a_3 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \hline a_{34} \\ \hline \end{array}$$

$$(3.1.30b) \quad y^+ = (X_1^+ : X_2^+) \begin{pmatrix} L \\ \delta \end{pmatrix} + a$$

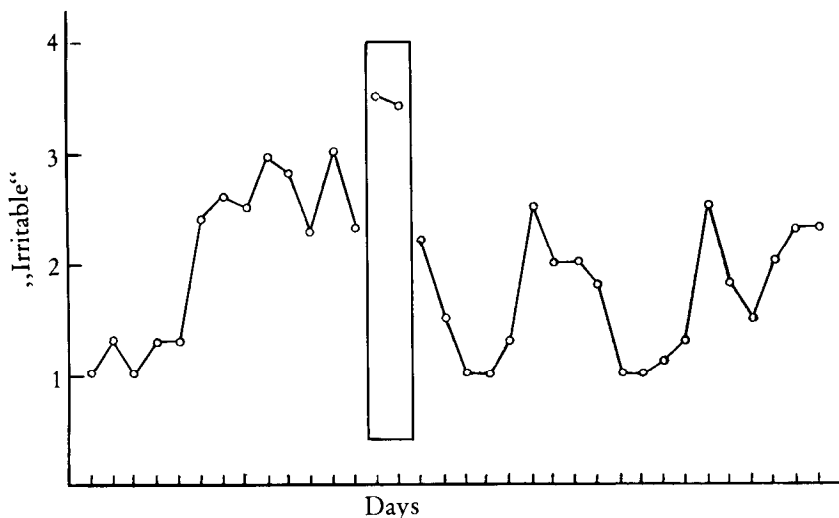


Fig. 3.1.5: Selbsteinschätzung der „Irritierbarkeit“ einer Frau vor, während und nach dem Menstruationsbeginn

Als Schätzungen für die Parameter geben Gottman Glass (1978) an:  $\hat{L} = 1.04$ ,  $\hat{\delta} = 1.21$ ,  $\hat{\theta}_1 = .14$ . Der t-Wert ist  $t_{32} = F_{1,32} = 3.34$  und somit signifikant. Andere Designmatrizen mit bis zu 5 Spalten in der Designmatrix X finden sich bei Meier (1981).

3.1.3 Verteilungsgebundene Prüfverfahren: Interventionsanalyse mit dem Transfermodell von Box & Tiao (1975)

Während der Ansatz von Box & Tiao (1965) und Glass et al. (1975) stark am allgemeinen linearen Modell orientiert war, haben Box & Tiao (1975) die Interventionskomponente konsequent in das Zeitreihenmodell integriert.

Das Modell kennt drei Variablensätze: (a) die Interventionsvariable  $I_t$ , die zu bestimmten a priori festgelegten Zeitpunkten die Werte 0 und 1 annimmt, je nachdem ein Interventionseffekt „ein“ oder „aus“ geschaltet wird, (b) die Effektivvariable  $Y_t^*$ , die eine Folge dieser Intervention ist und (c) die beobachtbare Zeitreihe  $Y_t$ , die einerseits durch die Effektivvariable  $Y_t^*$  andererseits durch anderweitige - nicht kontrollierte - Einflüsse geprägt wird:

$$Y_t = \left\{ \begin{matrix} \text{Interventionseffekte} \\ Y_t^* \end{matrix} \right\} + \left\{ \begin{matrix} \text{nichtkontrollierte Effekte} \\ N_t \end{matrix} \right\}$$

$Y^*$  stellt gewissermaßen das Niveau der Zeitreihe dar, das durch Handlungen des Experimentators hergestellt werden kann. Dieses durch Interventionen hergestellte Niveau, wird dann durch andere weitere Einflüsse, die nicht unter Kontrolle stehen, überlagert. Diese unkontrollierten Einflüsse werden in  $N_t$  zusammengefaßt und lassen sich durch ein  $ARIMA(p,d,q) (P,D,Q)_s$ -Modell repräsentieren.

Bei der Interventionsanalyse besteht die Aufgabe,  $Y_t$  wieder in  $Y_t^*$  und  $N_t$  zu zerlegen und die Parameter des Prozesses  $Y_t^*$  zu schätzen. Diese Parameter können als Effektparameter interpretiert und auf Signifikanz geprüft werden. Hätte Hilgard (1933) in einem Experiment zur Zwillingforschung (s. Figur 3.1.6) schon das Methodeninstrumentarium der Interventionsanalyse gekannt und verwendet, hätte er die Zeitreihen  $Y_{tC}$ ,  $Y_{tT}$  (durchgezogen) zerlegen müssen in einen Interventionsanteil  $Y_t^*$  und die nicht durch Intervention zustande gekommene Komponente  $N_t$ . In diesem Fall wäre der natürliche Reife-prozeß jedes Kindes (punktierte Linie) als deterministischer Trend in dem  $ARIMA(p,d,q)(P,D,Q)_s$ -Modell der  $N_t$ -Komponente zu interpretieren. Der Prozeß für  $N_t$  sollte mit den Preinterventionsdaten geschätzt werden (S. 305).

Betrachten wir zunächst das Zeitreihenmodell des Interventionseffektes. In einem ABAB-Design, das sich über 20 Tage erstreckt, wird die Interventionsvariable zweimal eingeschaltet („Stufenimpuls“):

t =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	A	A	A	A	A	B	B	B	B	B	A	A	A	A	A	B	B	B	B	B
$I_t =$	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1

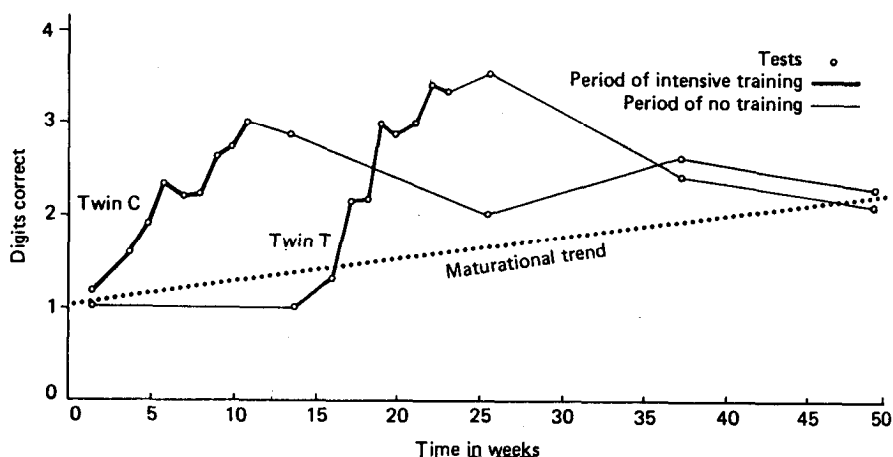


Fig. 3.1.6: Zeitreihenexperiment von Hilgard (1933) mit eineiigen Zwillingen und nichtkontrollierbarem Trend  $N_t$

Vermutet man, daß die Intervention eine abrupte Veränderung von  $Y_t$  (Sprung) ohne Zeitverzögerung hervorrufen würde und daß  $Y_t^*$  nur von einem ganz bestimmten Zeitindex  $I_{t-b}$  abhängt, verwenden wir die Transferfunktion 0. Ordnung (Figur 3.1.7a), wenn keine Zeitverzögerung des Interventionseffektes vorliegt (S. 306f.).

$$(3.1.32) \quad Y_t^* = \omega_0 I_t$$

oder wenn sich die Intervention  $b$  Zeitpunkte später auswirkt (Figur 3.1.7b):

$$(3.1.33) \quad Y_t^* = \omega_0 I_{t-b}$$

Ist die  $N_t$ -Komponente nichtstationär, verläuft das Niveau der Zeitreihe natürlich anders. So haben wir unter der Annahme, daß  $N_t$  einem linearen deterministischen Trend folgt, für die Niveaus der Zeitreihe  $E(Y_t | t)$  den Verlauf in Figur 3.1.7c.

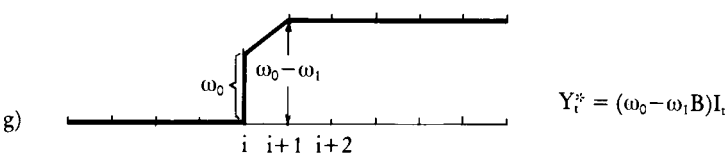
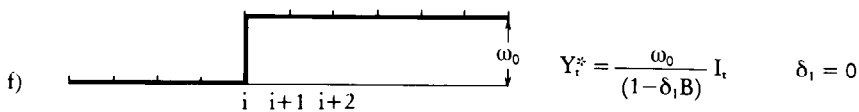
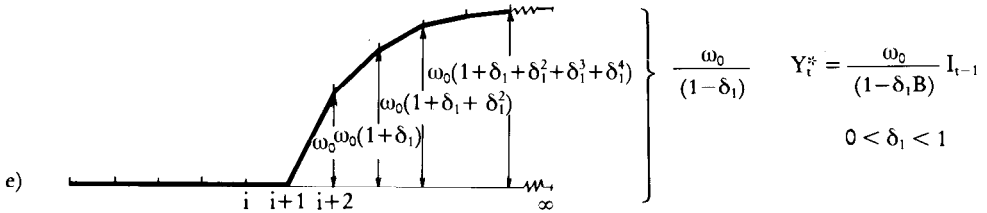
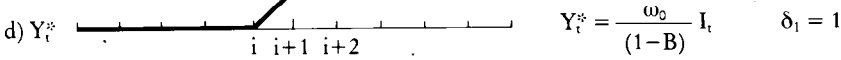
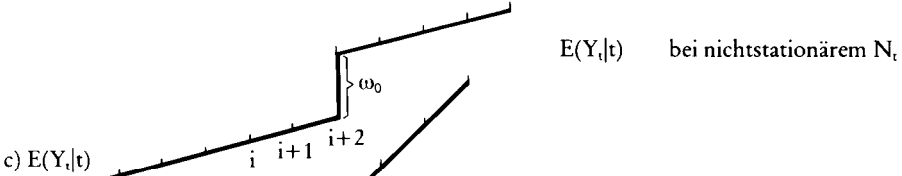
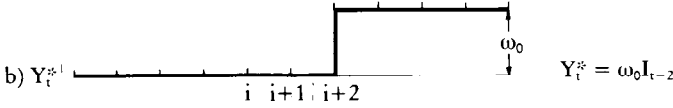
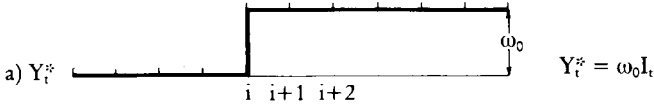
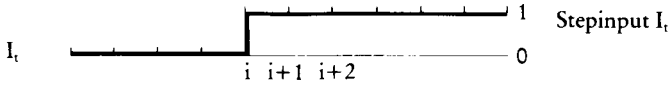
Tritt der Effekt auf  $Y_t^*$  vermutlich nicht plötzlich ein, sondern steigt nur allmählich an, benötigt man die Transferfunktion 1. Ordnung (Figur 3.7d,e). Liegt keine Verzögerung der Intervention vor, ist diese:

$$(3.1.34a) \quad Y_t^* = \delta_1 Y_{t-1}^* + \omega_0 I_t \text{ oder } (1 - \delta_1 B) Y_t^* = \omega_0 I_t$$

oder

$$(3.1.34b) \quad Y_t^* = \frac{\omega_0}{(1 - \delta_1 B)} I_t$$





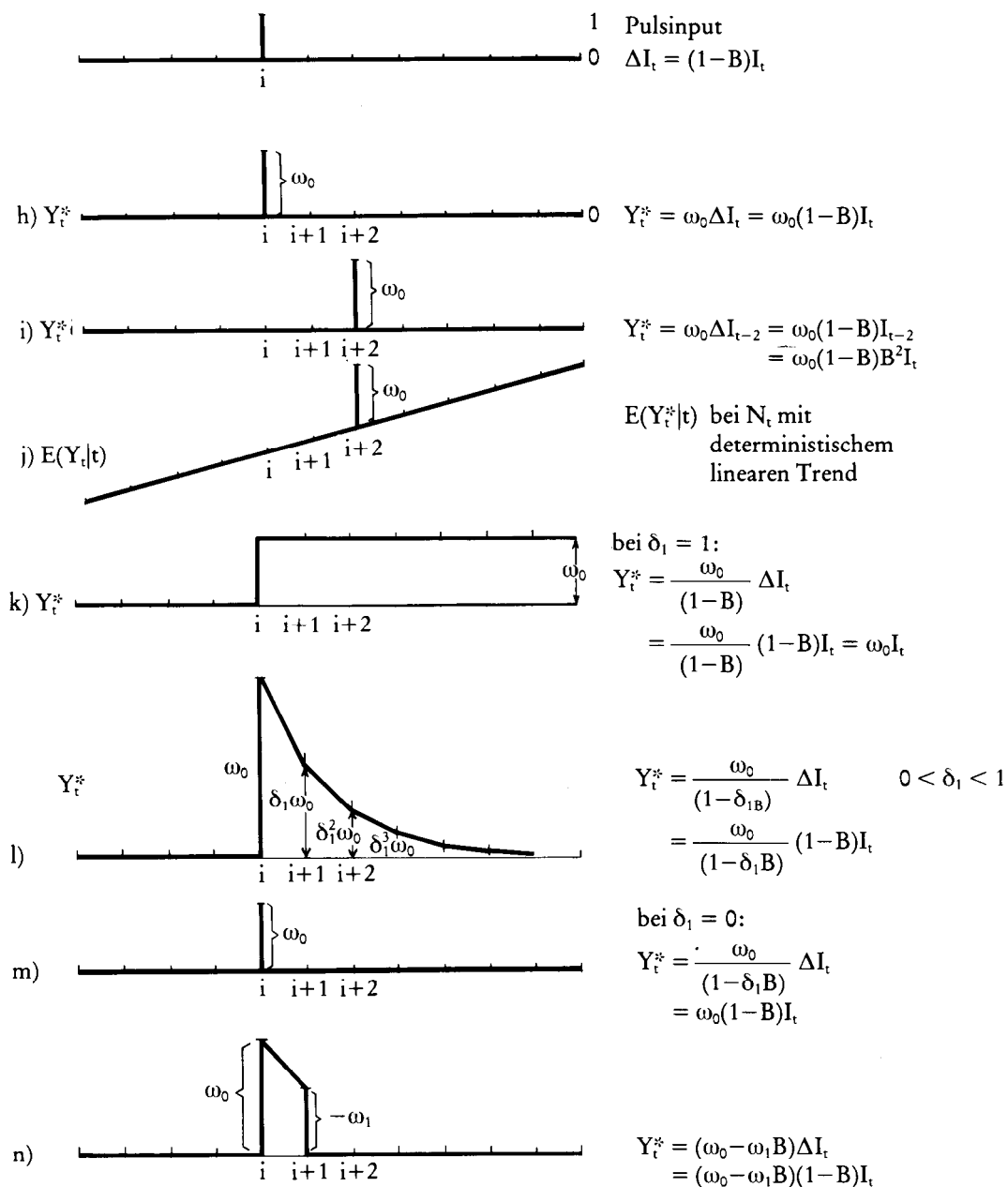


Fig. 3.1.7a-n: Verläufe der abhängigen Zeitreihe, die nur durch die Interventionen „Stepinput“ und „Pulsinput“ erzwungen werden. Diese deterministischen Verläufe werden von nichtkontrollierbaren stochastischen Prozessen überlagert.

bzw. mit Verzögerung des Interventionseffektes:

$$(3.1.35) \quad Y_t^* = \frac{\omega_0}{(1 - \delta_1 B)} I_{t-b} \quad (\text{s. Fig. 3.1.7e})$$

Bei  $\delta_1 = 0$  reduziert sich die Transferfunktion 1. Ordnung auf die der nullten Ordnung (Sprung von  $Y_t^*$  um 1 (s. Figur 3.1.7f und a). Ist dagegen  $\delta_1 = 1$ , haben wir

$$(3.1.36a) \quad (1 - B)Y_t^* = \omega_0 I_t$$

oder

$$(3.1.36b) \quad Y_t^* = Y_{t-1}^* + \omega_0 I_t$$

und damit einen nicht gedämpften Anstieg (Figur 3.1.7d). Die Stabilitätsgrenzen des Transfermodells sind daher  $-1 < \delta_1 < +1$ . Da das Transfermodell (3.1.34) eine Differenzengleichung 1. Ordnung ist, berechnet sich  $Y_t^*$  für die Interventionsperiode (alle  $I_t = 1$ ) als Lösung der Differenzengleichung (3.1.34) zu:

$$Y_{i+n}^* = \sum_{k=0}^n \delta_1^k \omega_0 = \omega_0 \sum_{k=0}^n \delta_1^k \quad \begin{array}{l} i = \text{Interventionsbeginn} \\ I_t = 1 \text{ für } t \geq i \end{array}$$

mit asymptotischem Niveau (s.a. Figur 3.1.7e)

$$Y_\infty^* = \omega_0 \sum_{k=0}^{\infty} \delta_1^k = \frac{\omega_0}{1 - \delta_1} \quad \text{für } |\delta_1| < 1$$

Hat man dagegen die Hypothese, daß  $Y_t^*$  nach dem Beginn der Intervention plötzlich anschnellt, um dann mehr oder minder schnell abzufallen, kann man die Transferfunktion 1. Ordnung beibehalten. Jedoch ist jetzt die Interventionsvariable  $I_t$  zu differenzieren. Das bedeutet inhaltlich:  $Y_t^*$  reagiert jetzt nur noch auf *Veränderungen* der Interventionsvariablen  $I_t$ . Die Differenzierung liefert  $(1 - B)I_t$ , der nicht mehr ein „Stufen-“ sondern ein „Pulsimpuls“ ist:

A	A	A	A	A	B	B	B	B	B
0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	1	0	0	0	0

$I_t$  „Stufenimpuls“

$(1 - B)I_t = I_t - I_{t-1}$  „Pulsimpuls“

$(1 - B)I_t$  nimmt nur zu Beginn einer Intervention den Wert 1 an. Es lassen sich jetzt eine Reihe von hypothetischen Effektverläufen  $Y_t^*$  darstellen (s. Figur 3.1.7h-n).

Ist im Modell

$$(3.1.37a) \quad (1 - \delta_1 B) Y_t^* = \omega_0 (1 - B) I_t$$

oder

$$(3.1.37b) \quad Y_t^* = \frac{\omega_0}{(1 - \delta_1 B)} (1 - B) I_t$$

$\delta_1 = 1$ , kürzen sich die Operatoren  $(1 - B)$  weg und es ist  $Y_t^* = \omega_0 I_t$ . Man erhält dann wieder die Transferfunktion 0. Ordnung (= abrupter Sprung) (Figur 3.1.7k).

Das allgemeine Transfermodell

$$(3.1.38a) \quad Y_t^* - \delta_1 Y_{t-1}^* - \delta_2 Y_{t-2}^* - \dots - \delta_r Y_{t-r}^* = \omega_0 I_{t-b} - \omega_1 I_{t-b-1} - \dots - \omega_s I_{t-b-s}$$

läßt sich mit Hilfe der Backshiftoperatoren verkürzt schreiben als

$$(3.1.38b) \quad (1 - \delta_1 B - \dots - \delta_r B^r) Y_t^* = (\omega_0 - \omega_1 B - \dots - \omega_s B^s) I_{t-b}$$

oder

$$(3.1.38c) \quad \delta(B) Y_t^* = \omega(B) I_{t-b}$$

oder

$$(3.1.38d) \quad Y_t^* = \frac{\omega(B)}{\delta(B)} I_{t-b} = \frac{\Omega(B)}{\delta(B)} I_t$$

Dabei spiegeln die Indices  $r, s$  das „Gedächtnis“ der Interventionskomponente wider.

Das Transferfunktionsmodell ist stabil (d.h. explodiert nicht), wenn die Beträge der reellen oder komplexen Wurzeln  $B_j$  des Polynoms

$$(3.1.39) \quad (1 - \delta_1 B - \dots - \delta_r B^r) = 0$$

größer als 1 sind. Die Stabilitätsgrenzen entsprechen numerisch den Grenzen der Stationarität von autoregressiven Prozessen. Für die Transferfunktion 2. Ordnung

$$(3.1.40) \quad (1 - \delta_1 B - \delta_2 B^2) Y_t^* = \omega_0 I_t$$

müssen die Parameter in folgenden Grenzen liegen:

$$\begin{aligned} -1 &< \delta_2 < +1 \\ \delta_1 + \delta_2 &< +1 \\ \delta_2 - \delta_1 &< +1 \end{aligned}$$

Nur dann sind die Wurzeln  $B_j$  des Polynoms  $(1 - \delta_1 B - \delta_2 B^2) = 0$  vom Betrag  $|B_j| > 1$ , wobei sich die Wurzeln berechnen nach

$$B_j = \frac{\delta_1 \pm \sqrt{\delta_1^2 + 4 \delta_2}}{2 \delta_2}$$

Hält man die Interventionsvariable 1, nach dem Zeitpunkt  $t=i$  konstant auf dem Wert 1 (Stepimpuls), ist für die Interventionskomponente  $Y_t^*$  folgendes Niveau zu erwarten:

$$(3.1.41) \quad Y_\infty^* = \frac{(\omega_0 - \omega_1 - \dots - \omega_s)}{(1 - \delta_1 - \dots - \delta_r)} = \frac{\omega(B) \cdot 1}{\delta(B) \cdot 1}$$

Liegen  $k$  Interventionsvariable gleichzeitig vor, kann man das Modell für die erzwungene Zeitreihe (3.1.38) im Sinne einer multifaktoriellen Varianzanalyse bzw. des multiplen Transfermodells (2.65) zum multiplen Interventionsmodell umformulieren:

$$(3.1.42) \quad Y_t^* = \sum_{j=1}^k Y_{jt}^* = \sum_{j=1}^k \left( \frac{\Omega_j(B)}{\delta_j(B)} \right) I_{jt}$$

So kann man sich in einem Wahrnehmungsexperiment, das den Effekt eines Tranquilizers auf die Wahrnehmungsgeschwindigkeit messen soll, den experimentellen Effekt  $Y_t^*$  aus zwei Komponenten zusammengesetzt denken:

$$(3.1.43) \quad Y_t^* = Y_{1t}^* + Y_{2t}^* = f(I_{1t}) + g(I_{2t})$$

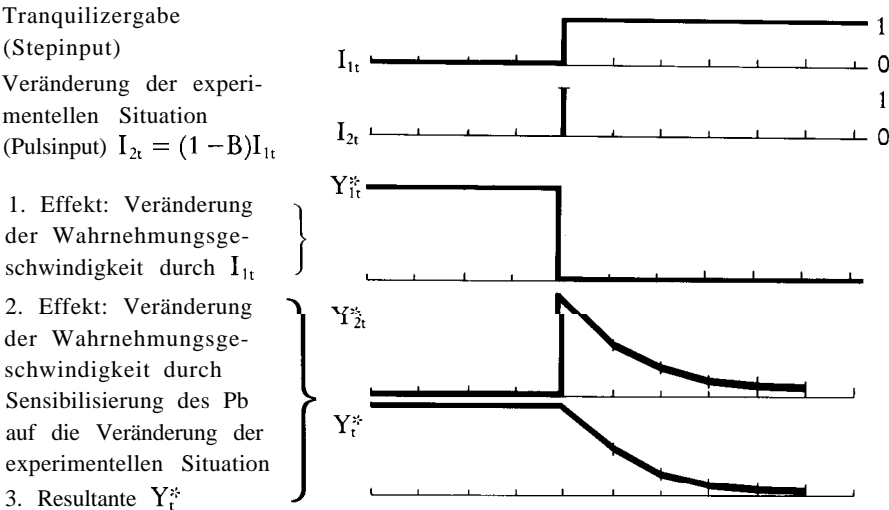


Fig. 3.1.8: Zusammensetzung des deterministischen Teils der Zeitreihe  $Y_t$

Als Ergebnis der experimentellen Faktoren erwarten wir einen langsamen Abfall der Wahrnehmungsgeschwindigkeit. Zu dem determinierten Verlauf  $Y_t^*$  addiert sich dann die Zeitreihe  $N_t$  zur beobachteten Zeitreihe  $Y_t$ :

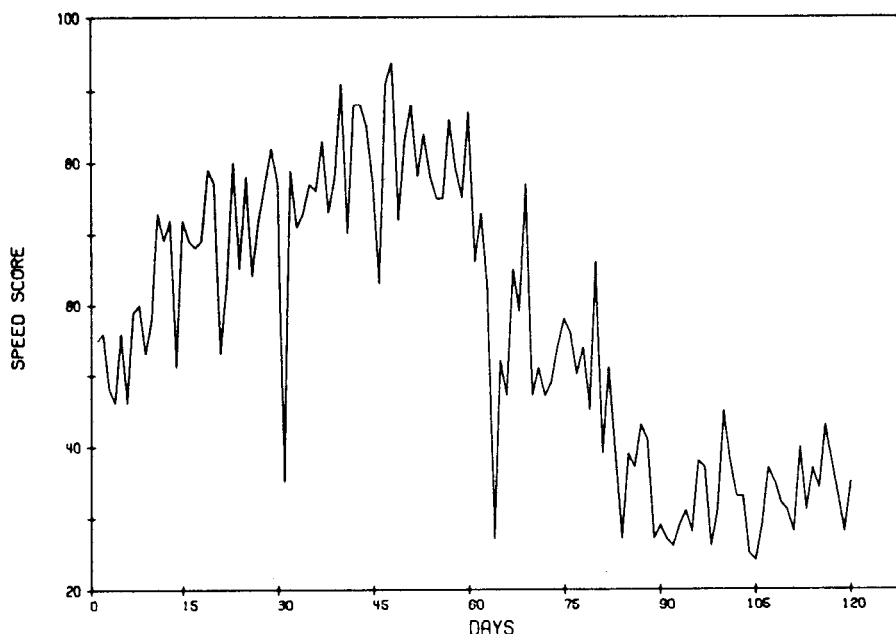


Fig. 3.1.9: Wahrnehmungsgeschwindigkeit vor und während der Medikamentengabe (Interventionsbeginn  $i=61$ )

Mit Hilfe der preexperimentellen Daten läßt sich der  $N_t$ -Prozeß schätzen (McCleary & Hay, 1980):

$$N_t = \frac{(1 - .77B)}{(1 - B)} a_t \quad \text{bzw.:} \quad \begin{aligned} N_t - N_{t-1} &= a_t - .77a_{t-1} \\ N_t &= N_{t-1} + a_t - .77a_{t-1} \end{aligned}$$

so daß das vollständige Interventionsmodell folgende Form annimmt

$$Y_t = Y_{1t}^* + Y_{2t}^* + N_t = \omega_{10}I_{1t} + \frac{\omega_{20}}{(1 - \delta_{21}B)} I_{2t} + N_t$$

mit Schätzungen:

$$Y_t = -28.8 * Y_t + \frac{18.05}{(1 - .51B)} (1 - B) I_t + \underbrace{\frac{(1 - .77 * B)}{(1 - B)}}_{N_t} a_t \quad I_t = \begin{cases} 0 & t < i \\ 1 & i \geq t \end{cases}$$

\*signifikant

In *Feldstudien* hat man natürlich die exogene Variable  $I_t$  nicht unter direkter Kontrolle. So kann man nur Vermutungen anstellen, ob z.B. eine Katastrophe oder Ölpreiserhöhung eine plötzliche oder eine allmähliche Änderung in  $Y_t^*$  erzwingt. Beobachten kann man nur  $Y_t$ .  $I_t$  bleibt im Gegensatz zum *Experiment* unbekannt. Für solche Ex-Post-Datenanalyse, die dem Vorgehen eines Historikers beim Aktenstudium nicht unähnlich ist, empfiehlt es sich folgende Modellfolge zu schätzen und deren Parameterzahl sukzessive zu reduzieren (Tabelle 3.1.2):

Schätzung des ARIMA (p, d, q) (P, D, Q) <sub>s</sub> -Modells von $N_t$ mit den Präinterventionsdaten			
Schätzung von $Y_t = \frac{\omega_0}{(1-\delta_1 B)} (1-B) I_{t-b} + N_t$ ; dabei werden die $N_t$ -Parameter festgehalten			
ja: permanenter Effekt (s. Fig. 3.1.7k)		nein: flüchtiger Effekt	
Schätzung von $Y_t = \frac{\omega_0}{(1-\delta_1 B)} I_{t-b} + N_t$		ja: Impulsantwort (s. Fig. 3.1.7m)	
		nein: allmählich verschwindender Effekt (s. Fig. 3.1.7l)	
ja: plötzlich eintretender Effekt (s. Fig. 3.1.7f)		nein: allmählich eintretender Effekt (s. Fig. 3.1.7e)	
Schätzung von $Y_t = \omega_0 I_{t-b} + N_t$ (s. Fig. 3.1.7a, b)	Es bleibt beim Modell $Y_t = \frac{\omega_0}{(1-\delta_1 B)} I_{t-b} + N_t$	Schätzung von $Y_t = \omega_0 (1-B) I_{t-b} + N_t$	Es bleibt beim Modell $Y_t = \frac{\omega_0 (1-B)}{(1-\delta_1 B)} I_{t-b} + N_t$

Ein Beispiel für eine Feldstudie mit einer hypothetisch angenommenen doppelten Interventionsvariablen bezieht sich auf eine Untersuchung von Hibbs (1977) zum Einfluß von Regierungswechsel (1. Interventionsvariable) und veränderter Sozialgesetzgebung (2. Interventionsvariable) auf die Zeitreihe  $Y_t$  der Arbeitslosigkeit in England.

Die Zeitreihe der prozentualen Arbeitslosigkeit findet sich in Figur 3.1.12. Die nichtstationäre  $N_t$ -Komponente wurde mit einem  $ARIMA(1,0,0)(0,1,1)$ -Prozeß gefittet (s.a. 2.25a).

$$(3.1.44) \quad N_t = \frac{\theta_0 + a_t}{(1 - B^4)(1 - \varphi_1 B)}$$

Das vollständige Transfermodell sah folgendermaßen aus:

$$(3.1.45) \quad Y_t = \frac{\omega_{10}}{(1 - \delta_1 B)} G_{t-1} + \frac{\omega_{20}}{(1 - \delta_2 B)} C_t + N_t$$

$Y_t$  = Prozentsatz der Arbeitslosen pro Quartal

$G_t$  = Interventionsvariable  $I_{1t}$  mit den Werten

$C_t$  = Interventionsvariable  $I_{2t}$  mit den Werten

$\left\{ \begin{array}{l} -1: \text{Labourregierung} \\ +1: \text{konserv. Regierung} \\ 0: \text{Zeit vor 1966} \\ 1: \text{Zeit nach 1965 mit} \\ \text{veränderter Sozial-} \\ \text{gesetzgebung} \end{array} \right.$

Hibbs nahm folgenden Verlauf für  $Y_{2t}^* = g(I_{2t})$  an (Figur 3.1.10 und 3.1.7e)

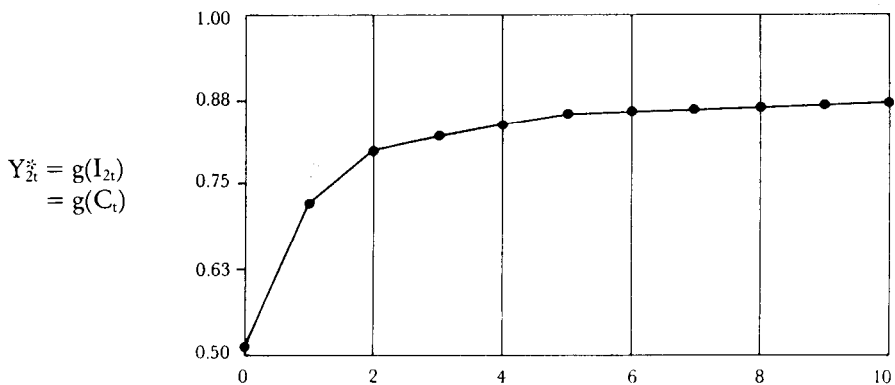


Fig. 3.1.10: Erwartete Niveauveränderung der Zeitreihe  $Y_{2t}^*$ , die nur durch die veränderte Sozialgesetzgebung ab 1965 bestimmt wäre

und folgenden Verlauf für  $Y_{1t}^*$  (Figur 3.1.11 und 3.1.7e).



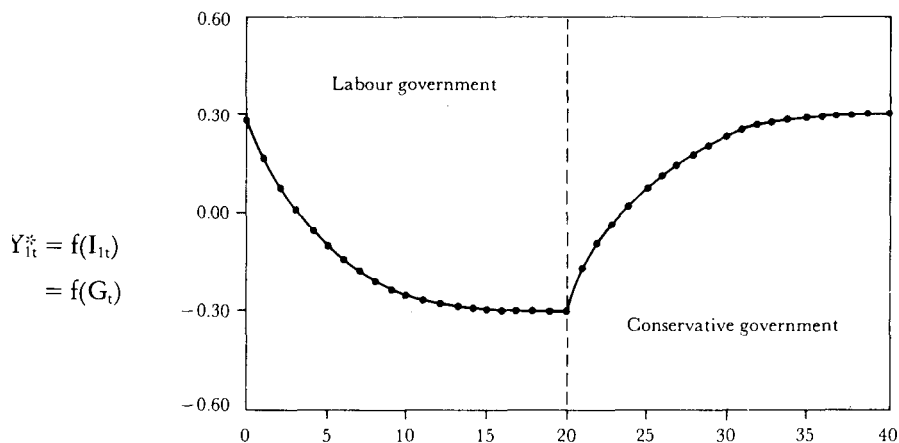


Fig. 3.1.1 : Erwartete Niveauveränderung der Zeitreihe  $Y_{1t}^*$ , die nur durch den Regierungswechsel bestimmt wäre

Ein Vergleich mit Figur 3.1.12 zeigt die gute Datenanpassung des Modells

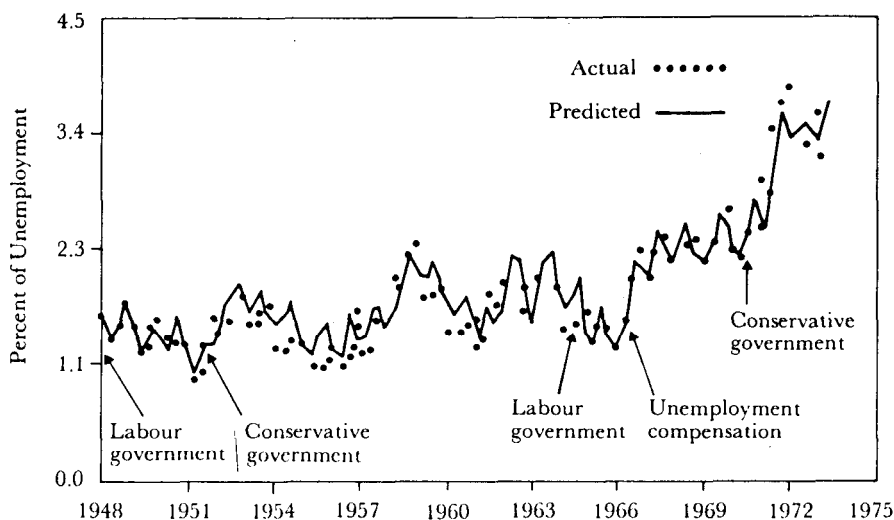


Fig. 3.1.12: Tatsächliche und mit Transfermodell vorhergesagte Arbeitslosenquote in England zwischen 1948 und 1973

Mit den Parametern des Transfermodells lassen sich Effekte isolieren, die im „Zick-Zack“ der Zeitreihe nicht zu erkennen sind.

Würde z.B.  $I_1$  konstant auf -1 gesetzt (Labour immer an der Regierung), müßte das Niveau der Zeitreihe  $Y_t^*$  um  $\omega_0/(1-\delta_1)$  sinken. Beim Wechsel der

Regierungsmacht auf die Konservativen wäre ein Anstieg um den gleichen Betrag zu erwarten, wenn alle anderen Variablen sich nicht ändern würden (Figur 3.1.11). Der isolierte Einfluß der Gesetzgebung auf die Arbeitslosenquote würde sich auf lange Sicht mit einem Anstieg um  $\omega_{20}/(1-\delta_2)$  auf das Niveau  $Y_t^*$  der Zeitreihe auswirken (s.a. Figur 3.1.10).

Allgemein lassen sich Transfer- und  $N_t$ -Modelle folgendermaßen kombinieren: gilt für den Interventionseffekt z.B. das Transfermodell

$$(3.1.46) \quad Y_t^* = (\omega_0 - \omega_1 B)I_t$$

und für das Modell  $N_t$ , das keine Interventionseffekte enthält, ein ARIMA  $(0,1,1)(0,1,1)_{12}$ :

$$(3.1.47a) \quad (1 - B)(1 - B^{12})N_t = (1 - \theta_1 B)(1 - \theta_{12} B^{12})a_t$$

bzw.

$$(3.1.47b) \quad N_t = \frac{(1 - \theta_1 B)(1 - \theta_{12} B^{12})}{(1 - B)(1 - B^{12})} a_t$$

ergibt die Kombination von (3.1.46) und (3.1.47) das vollständige Interventionsmodell

$$(3.1.48) \quad Y_t = (\omega_0 - \omega_1 B)I_t + \frac{(1 - \theta_1 B)(1 - \theta_{12} B^{12})}{(1 - B)(1 - B^{12})} a_t$$

Auf Querverbindungen zwischen den Modellen in Kapitel 3.1.2 und 3.1.3 weisen Möbus, Görücke & Kröh (1982) hin.

## 3.2 $N > 1$ Quasiexperimentelle Zeitreihendesigns (univariater Fall für eine Gruppe)

### 3.2.1 $N > T$ , $M = I$ , $G = I$

Stehen für jeden Meßzeitpunkt mehrere Personen zur Verfügung, liegt eine Zeitreihe von Querschnitten vor. Das quasiexperimentelle Zeitreihendesign mit  $N > 1$  läßt sich damit einordnen in die „Profilanalyse“ oder die „Wachstumskurvenanalyse“ einer Gruppe (s. a. Kap. 5). Auswertungsvorschläge hierzu legen (Algina & Swaminathan, 1977, 1979a, 1979b; Grizzle & Allen, 1969; Khatri, 1966; Marmor & Marmor, 1978; Morrison, 1967; Potthoff & Roy, 1964; Rao, 1959, 1965, 1966, 1967; Simonton, 1977; Swaminathan & Algina, 1977; Timm, 1975) vor.

Wird die Folge der Mittelwerte auf Interventionseffekte überprüft, sind u.a. zwei Nullhypothesen denkbar, die das Fehlen eines solchen Effektes postulieren :

$H_{01}$ : „alle Mittelwerte liegen auf einer *waagrecht*en Linie“

$H_{02}$ : „alle Mittelwerte liegen auf einer *geraden* Linie“

Die Überprüfung kann im Rahmen des allgemeinen linearen Modells (5.42) erfolgen:

$$(3.2.1) \quad {}_N Y_T = {}_N X_G B_T + {}_N E_T$$

wobei:  $N$  =Zahl der Personen

$G$  =Zahl der Gruppen (in 5.42 und bei Timm mit  $I$  bezeichnet)

$T$  = Zahl der Zeitpunkte (in 5.42 und bei Timm mit  $q$  geschrieben)

$Y$  = Rohdatenmatrix der abhängigen Variablen

$X$ = Designmatrix für das Design „zwischen den Gruppen“. Da wir hier nur eine Gruppe betrachten, ist  $X$  ein Spaltenvektor mit Komponenten gleich '1'.

$B$  = Parametermatrix mit Regressionskoeffizienten. Da wir hier nur *eine* Gruppe betrachten, ist  $B$  gleich einem Zeilenvektor  ${}_1 \mu'_T$  mit Populationsmittelwerten.

$E$  = Fehlermatrix

Jede Zeile von  $Y$  verteilt sich unabhängig normal mit Erwartungswertvektor  $\mu'$  und Kovarianzmatrix  $\Sigma$ , so daß

$$E({}_N Y_T) = {}_N X_G B_T = {}_N X_1 \mu'_T \text{ u n d } \text{cov}({}_N Y_T) = {}_N I_N \otimes {}_T \Sigma_T$$

wobei das Kroneckerprodukt z.B. bei Timm (1975, S. 30) erklärt ist.

Die Nullhypothesen  $H_{01}, H_{02}$  lassen sich als Matrixgleichung formulieren:

$$(3.2.2) \quad {}_1 C_1 B_T A_H = {}_1 \Gamma_H$$

wobei:  $C$  = Hypothesenmatrix für Vergleiche „zwischen den Gruppen“ bzw. den Zeilen der Parametermatrix  $B$  (hier ist  ${}_1 C_1 = 1$ )

$A$  = Hypothesenmatrix für die Vergleiche „*innerhalb* der Gruppen“ bzw. den Spalten der Parametermatrix oder „*zwischen* den Zeitpunkten“

$\Gamma$  = rechte Seiten der Hypothesengleichung. Meistens wird  $\Gamma = 0$  gesetzt.

Die Nullhypothese  $H_{01}$  „alle Mittelwerte liegen auf einer waagrecht en Linie“ läßt sich als Sammlung von  $H=(T-1)$  Einzelhypothesen

$$(3.2.3a) \quad \begin{aligned} \mu_1 - \mu_2 &= 0 \\ \mu_2 - \mu_3 &= 0 \\ &\vdots \\ \mu_{T-1} - \mu_T &= 0 \end{aligned}$$

oder nach (3.2.2) als Matrixgleichung

$${}_1\mu'_T A_H = {}_10_H$$

(3.2.3b) bzw.

$${}_1B_T A_H = {}_10_H$$

	1	0	0
	-1	1	:
	0	-1	:
	:	:	:
	:	:	:
	0	0	-1
$\mu_1 \mu_2 \dots \mu_T$	0	0	0

formulieren. Ähnlich verhält es sich mit der Nullhypothese  $H_{02}$  „alle Mittelwerte liegen auf einer *geraden* Linie“. Dieses bedeutet, daß die Differenzen benachbarter Mittelwerte einander gleich sind. Auch hier können die  $H=(T-2)$  Einzelhypothesen

$$(3.2.4a) \quad \begin{aligned} \mu_1 - \mu_2 &= \mu_2 - \mu_3 \\ \mu_2 - \mu_3 &= \mu_3 - \mu_4 \end{aligned}$$

$$\begin{aligned} \mu_1 - 2\mu_2 + \mu_3 &= 0 \\ \text{oder } \mu_2 - 2\mu_3 + \mu_4 &= 0 \end{aligned}$$

$$\mu_{T-2} - \mu_{T-1} = \mu_{T-1} - \mu_T$$

$$\mu_{T-2} - 2\mu_{T-1} + \mu_T = 0$$

nach (3.2.2) als Matrixgleichung geschrieben werden:

$${}_1\mu'_T A_H = {}_10_H$$

(3.2.4b) bzw.

$${}_1B_T A_H = {}_10_H$$

	1	0	0
	-2	1	:
	1	-2	:
	0	1	1
	:	:	-2
	:	:	1
$\mu_1 \mu_2 \dots \mu_T$	0	0	0

Zur Prüfung der Nullhypothesen kann man mehrere Kriterien verwenden. Die gebräuchlichsten sind Wilks Lambda und Hotelling's  $T^2$  (s. Timm, S. 230). Verwendet man Wilks  $\Lambda$ , muß die Nullhypothese verworfen werden, wenn

$$(3.2.5) \quad \Lambda = \frac{|Q_{\text{voll}}|}{|Q_{\text{restr}}|} < U_{\alpha, df_1 = H, df_2 = 1, df_3 = N - 1}$$

wobei  $|Q_{\text{voll}}|$  = Determinante der Kreuzproduktmatrix (= generalisierte Fehlervarianz) der Fehler des vollen - nicht eingeschränkten - Modells

$|Q_{\text{restr}}|$  = generalisierte Fehlervarianz des eingeschränkten Modells (Modell unter  $H_0$ )

$${}_H Q_{H^{\text{voll}}} = A'Y'[I - X(X'X)^{-1}X']YA = E'E$$

$${}_H Q_{H^{\text{restr}}} = Q_{\text{voll}} + A'\hat{B}'X'X\hat{B}A \text{ mit } \hat{B} = (X'X)^{-1}X'Y = {}_1\hat{\mu}_T$$

Die gleichen Hypothesen lassen sich auch mit Hotelling’s  $T^2$  prüfen. Die Nullhypothesen werden verworfen, wenn (Timm, S. 231)

(3.2.6)

$$T^2 = N_1 \bar{y}'_T A_H (A'_T S_T A)^{-1} A'_T \bar{y}_1 > T_{\alpha, df_1 = H, df_2 = N - 1}$$

wobei:  $\bar{y}'_T = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_T) =$  Vektor mit Stichprobenmittelwerten

$$S_T = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})(y_i - \bar{y})' = \text{Kovarianzmatrix der Zeitpunkte}$$

Wird die Nullhypothese (3.2.4) beibehalten, ist ein Interventionseffekt unwahrscheinlich. Es kann aber ein deterministischer linearer Trend vorliegen. Muß dagegen die Nullhypothese verworfen werden, kann das an einem nicht-linearen Trend oder an einem signifikanten Effekt liegen.

Nehmen wir z.B. folgendes Beispiel (Algina & Swaminathan, 1977). Das quasiexperimentelle Zeitreihendesign umfaßt zwei A- und B-Phasen. Kovarianzmatrix S und Mittelwertsvektor  $\bar{y}$  sind in Tabelle 3.2.1 aufgeführt.

Tabelle 3.2.1: Kovarianzmatrix und Mittelwertsvektor in einem Zeitreihendesign

	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>	A <sub>4</sub>	A <sub>5</sub>	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>	B <sub>4</sub>	B <sub>5</sub>
A <sub>1</sub>	1.210									
A <sub>2</sub>	.425	1.254								
A <sub>3</sub>	.849	.694	1.690							
A <sub>4</sub>	.599	.511	.873	1.822						
A <sub>5</sub>	.624	.252	.643	.514	.960					
B <sub>1</sub>	.492	.284	.522	.362	.236	2.560				
B <sub>2</sub>	.592	.275	.588	.403	.335	.696	1.440			
B <sub>3</sub>	.268	.174	.314	.303	.164	.746	.698	1.690		
B <sub>4</sub>	.466	.263	.496	.368	.218	.508	.737	.780	1.960	
B <sub>5</sub>	.264	.194	.240	.089	.092	.499	.735	.492	.505	1.440
$\bar{y}' =$	[10.2	10.6	10.9	11.5	11.9	13.2	14.2	14.9	15.9	16.8]

Die Nullhypothese „alle Mittelwerte liegen auf einer *geraden* Linie“ (3.2.4) führt zu einem Hotelling’s  $T^2$  (3.2.6)

$$T^2 = 74.25 > T_{0.01, df_1=8, df_2=39} = 33.9$$

Daher wird die Nullhypothese verworfen. Anschließend stellt sich die Aufgabe, genauere Hypothesen über den Mittelwertsverlauf aufzustellen und zu testen. Entsprechend einem Vorschlag von Potthoff & Roy (1964) wird das

allgemeine Modell (3.2.1) durch die Nachmultiplikation mit der Designmatrix „innerhalb der Personen“ bzw. „zwischen den Zeitpunkten“ verändert zu (s.a. 5.41)

$$(3.2.7) \quad {}_N Y_T = {}_N X_1 B_R P_T + {}_N E_T \quad (P \text{ wird in (5.41) und bei Timm mit } Q \text{ bezeichnet})$$

mit  $\text{cov}(Y) = {}_N I_N \otimes {}_T \Sigma_T$  ( $R$  gibt die Zahl von Trendkomponenten: konst., lin., quadr. . . . an)

Eine Schätzung für  $\Sigma$  ist  $S = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})(y_i - \bar{y})'$ . Es ist dabei zu beach-

ten, daß  $B$  in (3.2.7) bei  $R < T$  und  $P \neq I$  nicht mehr wie in (3.2.1) die Populationsmittelwerte enthält, sondern grundlegendere Basisparameter, die die Unterschiede zwischen den Populationsmittelwerten beschreiben. Die Matrix  $P$  kann dabei verschiedene Formen je nach spezifiziertem Zeitreihendesign annehmen. So kann  $P$  eine Vandermonde-Matrix (s. Timm, S. 500), eine Matrix mit Koeffizienten orthogonaler Polynome (s. Winer, 1971, S. 878) oder eine Matrix mit normierten orthogonalen Polynomen sein.

Wir wählen hierzu ein kleines Beispiel mit  $N=5$  und  $T=3$

$${}_N Y_T = \begin{bmatrix} 2 & 4 & 7 \\ 2 & 6 & 10 \\ 3 & 7 & 10 \\ 7 & 9 & 11 \\ 6 & 9 & 12 \end{bmatrix} \quad {}_N X_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad {}_1 B_R = [b_1 \ b_2 \ b_3]$$

$${}_1 \bar{y}_T = [4 \ 7 \ 10]$$

Ist  $P_1$  eine Vandermonde-Matrix

$${}_R P_{1T} = \left. \begin{array}{c} \text{Zeitpunkte} \rightarrow \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 4 & 9 \end{bmatrix} \end{array} \right\} \begin{array}{l} \text{konstanter} \\ \text{linearer} \\ \text{quadratischer} \end{array} \text{Trend}$$

ist der erwartete Mittelwertsvektor für die drei Zeitpunkte

$${}_1 E(Y)'_T = {}_1 B_{1R} P_{1T} = b_{11} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}' + b_{12} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}' + b_{13} \begin{pmatrix} 1 \\ 4 \\ 9 \end{pmatrix}'$$

Ist dagegen  $P_2$  eine Matrix mit Koeffizienten *orthogonaler Polynome*,

$${}_R P_{2T} = \left. \begin{array}{c} \text{Zeitpunkte} \rightarrow \\ \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & -2 & 1 \end{bmatrix} \end{array} \right\} \begin{array}{l} \text{konstanter} \\ \text{linearer} \\ \text{quadratischer} \end{array} \text{Trend}$$

so ist der erwartete Mittelwertsvektor für die drei Zeitpunkte

$${}_1E(Y)'_T = {}_1B_{2R}P_{2T} = b_{21} \begin{pmatrix} +1 \\ +1 \\ +1 \end{pmatrix}' + b_{22} \begin{pmatrix} -1 \\ 0 \\ +1 \end{pmatrix}' + b_{23} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}'$$

Ist  $P_3$  eine Matrix mit normierten Koeffizienten orthogonaler Polynome, so daß

$${}_R P_{3T} = \left[ \begin{array}{ccc} \text{Zeitpunkte} \rightarrow \\ \begin{matrix} .577350 & .577350 & .577350 \\ -.707107 & .000000 & .707107 \\ .408248 & -.816497 & .408248 \end{matrix} \end{array} \right] \left. \begin{array}{l} \text{konstanter} \\ \text{linearer} \\ \text{quadratischer} \end{array} \right\} \text{Trend}$$

ist der erwartete Mittelwertsvektor für die drei Zeitpunkte

$${}_1E(Y)'_T = {}_1B_{3R}P_{3T} = b_{31} \begin{pmatrix} +.5 \\ +.5 \\ +.5 \end{pmatrix}' + b_{32} \begin{pmatrix} -.7 \\ 0 \\ +.7 \end{pmatrix}' + b_{33} \begin{pmatrix} .4 \\ -.8 \\ .4 \end{pmatrix}'$$

Ist der Trend vollständig spezifiziert, ist der Grad des Polynoms (T-1). Dann ist  $R=T$  und die Matrix  $P$  quadratisch. In diesem Fall und wenn  $P$  invertierbar ist, gestaltet sich die Schätzung der Regressionskoeffizienten besonders einfach. Die Schätzungen lassen sich nach der Methode der kleinsten Quadrate berechnen:

$$(3.2.8a) \quad \text{für } P_1: \hat{B}_1 = (X'X)^{-1}X'YP_1^{-1} = \bar{y}'P_1^{-1} = [1.00 \ 3.00 \ 0.00]$$

$$(3.2.8b) \quad \text{für } P_2: \hat{B}_2 = (X'X)^{-1}X'YP_2^{-1} = \bar{y}'P_2^{-1} = [7.00 \ 3.00 \ 0.00]$$

$$(3.2.8c) \quad \text{für } P_3: \hat{B}_3 = (X'X)^{-1}X'YP_3' = \bar{y}'P_3' = [12.12 \ 4.24 \ 0.00]$$

Hypothesen über  $B$  werden dann wieder wie unter (3.2.2) geprüft mit

$$(3.2.9) \quad {}_1C_1B_RA_H = {}_1\Gamma_H$$

Die dazugehörigen generalisierten Fehlervarianzen bestimmen sich nach

$$(3.2.10a) \quad |Q_{\text{voll}}| = |A'P^{-1}Y'[I - X(X'X)^{-1}X']YP^{-1}A|$$

$$(3.2.10b) \quad |Q_{\text{restr}}| = |Q_{\text{voll}} + (C\hat{B}A)'[C(X'X)^{-1}C']^{-1}(C\hat{B}A)|$$

Die Nullhypothese: „es liegt kein linearer oder quadratischer ‘Trend vor‘“, führt zu folgender Hypothesengleichung (3.2.11) (S. 321).

Oft muß aber  $R < T$  angenommen werden: die Zahl der Effekte, die die Variation der Mittelwertsreihe beschreibt, ist kleiner als die Zahl der Zeitpunkte. Hierdurch fließen schon in die Formulierung der Designmatrix  $P$  Hypothesen ein, die auf ihre Berechtigung hin überprüft werden müssen. Dieser nur im Falle  $R < T$  vorzuschaltende Hypothesentest ist als „test of the fit of the model“ (Grizzle & Allen, 1969) bekannt geworden. Für diese und die

$$(3.2.11a) \quad {}_1C_1 B_3 A_2 = {}_1I_2$$

$$(3.2.11b) \quad \begin{array}{c} \text{konst.} \\ \text{lin.} \\ \text{quadr.} \end{array} \begin{array}{|c|} \hline 1 \\ \hline \end{array} \begin{array}{|c|} \hline b_1 \\ \hline \end{array} \begin{array}{|c|} \hline b_2 \\ \hline \end{array} \begin{array}{|c|} \hline b_3 \\ \hline \end{array} \begin{array}{|c|} \hline 0 \quad 0 \\ 1 \quad 0 \\ 0 \quad 1 \\ \hline \end{array} \begin{array}{l} \text{konst.} \\ \text{lin.} \\ \text{quadr.} \end{array} = \begin{array}{|c|} \hline 0 \quad 0 \\ \hline \end{array}$$

$b_1$  = Regressionskoeffizient für Konstante

$b_2$  = Regressionskoeffizient für linearen Trend

$b_3$  = Regressionskoeffizient für quadratischen Trend

weiteren Betrachtungen empfiehlt es sich, das Linearmodell (3.2.7) zu vereinfachen. Multipliziert man beide Seiten von (3.2.7) mit  $\frac{1}{N} X'$  von links

$$\frac{1}{N} ({}_1X'_N Y_T) = \frac{1}{N} ({}_1X'_N X_1 B_R P_T) + \frac{1}{N} ({}_1X'_N E_T)$$

erhält man

$$(3.2.12a) \quad {}_1\bar{Y}'_T = {}_1B_R P_T + {}_1\bar{e}'_T \quad \text{da:} \quad \frac{1}{N} X'X = 1$$

bzw. transponiert:

$$(3.2.12b) \quad {}_T\bar{Y}_1 = {}_T P'_R B'_1 + {}_T \bar{e}_1$$

ein Linearmodell in den Mittelwerten:

$$(3.2.12c) \quad \begin{array}{c} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ \hline B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \end{array} \begin{array}{|c|} \hline 10.2 \\ 10.6 \\ 10.9 \\ 11.5 \\ 11.9 \\ \hline 13.2 \\ 14.2 \\ 14.9 \\ 15.9 \\ 16.8 \\ \hline \end{array} = \begin{array}{|c|} \hline \\ \\ \\ \\ \\ \hline \end{array} {}_T P'_R \cdot \begin{array}{|c|} \hline b_1 \\ b_2 \\ \vdots \\ b_R \\ \hline \end{array} + \begin{array}{|c|} \hline \\ \\ \\ \\ \\ \hline \end{array} {}_T \bar{e}_1$$

Je nach der Form der Zeitdesignmatrix  $P'$  lassen sich jetzt verschiedene Mittelwertsverläufe modellieren:



a) das Modell einer permanenten Niveauveränderung:

	1	10.2		1	0	$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$	$+ \tau \bar{e}_1$
A-Phase	:	:		1	0		
	:	:		1	0		
	:	:		:	:		
	:	:		:	:		
	:	13.2	=	:	1		
B-Phase	:	:		:	1		
	:	:		1	1		
	:	:		1	1		
T	:	:		1	1		

b) das Modell einer vorübergehenden Niveauveränderung:

	1	10.2		1	0	$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$	$+ \tau \bar{e}_1$
A-Phase	:	:		1	0		
	:	:		:	:		
	:	:		1	0		
	:	:		:	:		
	:	13.2	=	1	1		
B-Phase	:	:		1	0		
	:	:		:	:		
	:	:		:	:		
T	:	:		1	0		

c) das Modell einer abklingenden Niveauveränderung

	10.2		1	0	$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$	$+ \bar{e}$
A-Phase	10.6		1	0		
	10.9		1	0		
	11.5		1	0		
	11.9		1	0		
		=				
	13.2		1	1		
B-Phase	14.2		1	.5		
	14.9		1	.25		
	15.9		1	.125		
	16.8		1	.0625		

d) das Modell mit höherem gemeinsamen Trend (lin., quadr., kub.)

A-Phase	10.2	1	-9	6	-42	$\begin{matrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{matrix}$	$+ \bar{e}$
	10.6	1	-7	2	14		
	10.9	1	-5	-1	35		
	11.5	1	-3	-3	31		
	11.9	1	-1	-4	12		
B-Phase	13.2	1	1	-4	-12		
	14.2	1	3	-3	-31		
	14.9	1	5	-1	-35		
	15.9	1	7	2	-14		
	16.8	1	9	6	42		

e) das Modell mit Trends und Niveauänderung (gemeins. lin. Trend)

A-Phase	10.2	1	-9	0	$\begin{matrix} b_1 \\ b_2 \\ b_3 \end{matrix}$	$+ \bar{e}$
	10.6	1	-7	0		
	10.9	1	-5	0		
	11.5	1	-3	0		
	11.9	1	-1	0		
B-Phase	13.2	1	1	1		
	14.2	1	3	.5		
	14.9	1	5	.25		
	15.9	1	7	.125		
	16.8	1	9	.0625		

f) das Modell mit Strukturbruch (veränderte Regressionen im Sinne einer neuen Konstanten und Steigung)

A-Phase	10.2	1	-2	0	0	$\begin{matrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{matrix}$	$+ \bar{e}$
	10.6	1	-1	0	0		
	10.9	1	0	0	0		
	11.5	1	1	0	0		
	11.9	1	2	0	0		
B-Phase	13.2	0	0	1	-2		
	14.2	0	0	1	-1		
	14.9	0	0	1	0		
	15.9	0	0	1	1		
	16.8	0	0	1	2		

g) das Modell mit veränderten Trends, so daß  $R = 6$  ist: bei Verwendung der Vandermode-Matrix (links) und orthogonaler Polynome (rechts)

$$\begin{array}{c}
 \begin{array}{c} \text{A-Phase} \\ \text{B-Phase} \end{array} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 4 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 1 & 4 & 16 & 0 & 0 & 0 \\ 1 & 5 & 25 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 2 & 4 \\ 0 & 0 & 0 & 1 & 3 & 9 \\ 0 & 0 & 0 & 1 & 4 & 16 \\ 0 & 0 & 0 & 1 & 5 & 25 \end{bmatrix} = P' \\
 \begin{array}{cc} 1 \dots r & 1 \dots r \end{array}
 \end{array}
 \quad
 \begin{array}{c}
 \begin{array}{c} \text{A-Phase} \\ \text{B-Phase} \end{array} \begin{bmatrix} 1 & -2 & 2 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -2 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 0 \\ 1 & 2 & 2 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & -2 & 2 \\ 0 & 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 1 & 2 & 2 \end{bmatrix} = P' \\
 \begin{array}{cc} 1 \dots r & 1 \dots r \end{array}
 \end{array}$$

Besitzen die A- und B-Phasen unterschiedliche Länge (z.B.:  $T_A=5$  und  $T_B=6$ ), dann kann das Modell mit veränderten nichtlinearen Trends folgendermaßen formuliert werden (h) (S. 325).

Sind die Trends für A- und B-Phasen voll spezifiziert, so daß  $R = T_A + T_B$  wird  $P$  quadratisch von der Ordnung  $T_A + T_B$  (s. Modell i auf S. 325).

In diesem Fall ist aber der Vektor  $\bar{e} = 0$ , weil sich die  $T_A$  Mittelwerte der A- und die  $T_B$  Mittelwerte der B-Phase perfekt durch Polynome vom Grade  $r_A=T_A-1$  und  $r_B=T_B-1$  abbilden lassen.

Der Mittelwertsvektor  $\bar{e}$  in a)-h) läßt sich als Mittelwertsvektor zusätzlicher im Design  $P$  „vergessener“ Variabler auffassen. Trifft das Linearmodell a)-h) zu, darf der Mittelwertsvektor  $\bar{e}$  der „fehlenden“ Variablen nicht signifikant von Null abweichen, da

$${}_T E(\bar{y})_1 = {}_T P'_R B_1 \quad \text{und} \quad {}_T E(\bar{e})_1 = {}_T 0_1$$

ist, wenn das Modell gilt.





Daher muß geprüft werden, ob der Mittelwertsvektor  $\bar{e}$  oder der linear transformierte Vektor  $H'\bar{e}$  signifikant von Null abweicht. Hierzu suchen wir eine nicht verschwindende Matrix  $H'$ , so daß

$${}_{T-R} H'_T P'_R = 0$$

ist. Dann transformieren wir (3.2.12b) zu

$$(3.2.13) \quad {}_{T-R} H' \tilde{y}_1 = {}_{T-R} H'_T P'_R B_1 + {}_{T-R} H'_T \bar{e}_1$$

- h) das Modell mit veränderten nichtlinearen Trends ( $R = 8 < T_A + T_B$ ) bei unterschiedlich langen A- und B-Phasen

	bei Verwendung der Vandermonde-Matrix		und bei orthogonalen Polynomen																																																	
A-Phase	<table><tr><td>1</td><td>1</td><td>1</td><td>1</td></tr><tr><td>1</td><td>2</td><td>4</td><td>8</td></tr><tr><td>1</td><td>3</td><td>9</td><td>27</td></tr><tr><td>1</td><td>4</td><td>16</td><td>64</td></tr><tr><td>1</td><td>5</td><td>25</td><td>125</td></tr></table> 	1	1	1	1	1	2	4	8	1	3	9	27	1	4	16	64	1	5	25	125		<table><tr><td>1</td><td>-2</td><td>2</td><td>-1</td></tr><tr><td>1</td><td>-1</td><td>-1</td><td>2</td></tr><tr><td>1</td><td>0</td><td>-2</td><td>0</td></tr><tr><td>1</td><td>1</td><td>-1</td><td>-2</td></tr><tr><td>1</td><td>2</td><td>2</td><td>1</td></tr></table> 	1	-2	2	-1	1	-1	-1	2	1	0	-2	0	1	1	-1	-2	1	2	2	1									
1	1	1	1																																																	
1	2	4	8																																																	
1	3	9	27																																																	
1	4	16	64																																																	
1	5	25	125																																																	
1	-2	2	-1																																																	
1	-1	-1	2																																																	
1	0	-2	0																																																	
1	1	-1	-2																																																	
1	2	2	1																																																	
B-Phase	 <table><tr><td>1</td><td>1</td><td>1</td><td>1</td></tr><tr><td>1</td><td>2</td><td>4</td><td>8</td></tr><tr><td>1</td><td>3</td><td>9</td><td>27</td></tr><tr><td>1</td><td>4</td><td>16</td><td>64</td></tr><tr><td>1</td><td>5</td><td>25</td><td>125</td></tr><tr><td>1</td><td>6</td><td>36</td><td>216</td></tr></table>	1	1	1	1	1	2	4	8	1	3	9	27	1	4	16	64	1	5	25	125	1	6	36	216		 <table><tr><td>1</td><td>-5</td><td>5</td><td>-5</td></tr><tr><td>1</td><td>-3</td><td>-1</td><td>7</td></tr><tr><td>1</td><td>-1</td><td>-4</td><td>4</td></tr><tr><td>1</td><td>1</td><td>-4</td><td>-4</td></tr><tr><td>1</td><td>3</td><td>-1</td><td>-7</td></tr><tr><td>1</td><td>5</td><td>5</td><td>5</td></tr></table>	1	-5	5	-5	1	-3	-1	7	1	-1	-4	4	1	1	-4	-4	1	3	-1	-7	1	5	5	5	
1	1	1	1																																																	
1	2	4	8																																																	
1	3	9	27																																																	
1	4	16	64																																																	
1	5	25	125																																																	
1	6	36	216																																																	
1	-5	5	-5																																																	
1	-3	-1	7																																																	
1	-1	-4	4																																																	
1	1	-4	-4																																																	
1	3	-1	-7																																																	
1	5	5	5																																																	

$P' = \begin{array}{|c|c|} \hline P'_1 & 0 \\ \hline 0 & P'_2 \\ \hline \end{array}$

- i) das Modell mit veränderten nichtlinearen Trends ( $R = 11 = T_A + T_B$ ) bei unterschiedlich langen A- und B-Phasen

A-Phase	1	1	1	1	1																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																														
---------	---	---	---	---	---	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Ist das Modell korrekt, ist unter  $H_0$

$$(3.2.14) \quad E(H'\bar{y}) = E(H'\bar{e}) = 0$$

$H'$  kann als Zeilenbasis von  $[T_R P'_R (P P')_R^{-1} P_T - T_I I_T]$  bestimmt werden (Grizzle & Allen, 1969, S. 366). So ist z.B. als eine Möglichkeit

$$(3.2.15) \quad T_{-R} H'_T = T_{-R} T'_T [P'_R (P_T P')_R^{-1} P_T - T_I I_T]$$

mit beliebiger Matrix  $T$  denkbar. Weitere Vorschläge finden sich bei Grizzle & Allen (1969), Swaminathan & Algina (1977), Timm (1975, S. 496). Besonders einfach läßt sich  $H$  finden, wenn  $P'$  eine Matrix mit orthogonalen Polynomkoeffizienten ist. Die Zeilen von  $H'$  bestehen dann aus „überflüssigen“  $(T - R)$  Polynomkoeffizienten, wie man an den schraffierten Teilen der Matrix  $P'$  in i) sieht, wenn man für A- und B-Phase verschiedene Polynome von gemeinsamem Grade (und damit  $R = 8$ ) annimmt.



für Modell f):  $H_0$  kein Strukturbruch  $b_1 = b_3$  und  $b_2 = b_4$

$$[1] \quad \begin{bmatrix} b_1 & b_2 & b_3 & b_4 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \end{bmatrix}$$

für Modell g):  $H_0$  kein Strukturbruch

$$[1] \quad \begin{bmatrix} b_1 & b_2 & b_3 & | & b_4 & b_5 & b_6 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

Die Prüfung der Hypothesen kann wieder über Wilks' Lambda erfolgen, wobei die Berechnung der generalisierten Fehlervarianzen wegen der Nichtinvertierbarkeit der nichtquadratischen Matrix  $P$  nach (5.47) erfolgt (dort ist für  $K'$  die Matrix  $P$  einzusetzen).

Die Nullhypothese wird verworfen, wenn

$$\Lambda = \frac{|Q_{\text{voll}}|}{|Q_{\text{restr}}|} < U_{\alpha, df_1=H, df_2=G, df_3=N-G-(T-R)}$$

ist. Wir werden im Kap. 7.2 zeigen, wie man (3.2.18) und einfache Hypothesen („Sind Parameter Null oder einander gleich?“) mit LISREL prüfen kann.

### 3.2.2 $G > 1$ Quasiexperimentelle Zeitreihendesigns bei mehreren Gruppen

Hierzu erhält die Designmatrix  $X$  und die Matrix  $B$  eine Erweiterung auf  $G$ -Spalten bzw.  $G$ -Zeilen, wie es in (3.2.1) schon angedeutet ist. In (3.2.8) wird der Mittelwertsvektor  $\bar{y}_T$  durch eine Mittelwertsmatrix  ${}_G\bar{Y}_T$  ersetzt. Siehe auch Näheres im Kapitel 5 über Varianzanalyse.

### 3.2.3 $M > 1$ Quasiexperimentelle Zeitreihendesigns mit mehreren abhängigen Variablen

Bei  $M > 1$  abhängigen Variablen wird die Matrix  $P$  in (3.2.7) erweitert zu

		1 ... T		1 ... T	
1 : R		$P$	0	0	
		0			0
		1 : M : R		0	$P$
		1 ..... M			

die Matrix  $Y$  in (3.2.7) zu:

		1 ... T		1 ... T	
1 : N		$Y_1$	.....	$Y_M$	

und die Matrix (bzw. Vektor)  $B$  in (3.2.7) zu

		1 ... R		1 ... R	
1 : G		$B_1$	.....	$B_M$	

Die Testkriterien verändern sich dazu analog. Näheres siehe auch im Kapitel 5 über Varianzanalyse und Timm (1975, S. 507f.).

4. Veränderungsmessung mit Hilfe von Differenzenwerten

$N > 1, \quad T = 2$

Werden zu zwei Zeitpunkten  $t_1$  (Pretest) und  $t_2$  (Posttest) Messungen  $Y_1$  und  $Y_2$  erhoben, kann man den Differenzwert  $D = Y_2 - Y_1$  als ein Maß für Veränderung ansehen. Dieses Pretest-Posttestkonzept spielt z.B. in der Lernfähigkeitsdiagnostik bei Lerntests (Guthke 1976, 1977, 1980; Clauss, Guthke & Lehwald, 1978; Melchinger, 1978; Möbus, 1981) und in der Hirnschadendiagnostik (Möbus & Wallasch, 1977; Becker & Schmidtke, 1977) eine Rolle. Die methodischen Überlegungen sollen an einem inhaltlich zwar trivialen, dafür

aber einfachen Beispiel („Gewichtsveränderung“ durch eine Erholungskur) diskutiert werden.

Vorerst soll der Differenzwert  $D = Y_2 - Y_1$  (Dabei sei  $Y_2$  der Endwert (z.B.: Gewicht nach Erholungskur) und  $Y_1$  der Anfangswert (z.B.: Gewicht vor Kur)) betrachtet werden. Die Konzeption für diesen Veränderungsindex ist sehr einfach. Die Anfangs- und Endwerte müssen nur Intervallskalen aufweisen, damit die Differenzbildung sinnvoll ist. Allseits bekannt sind diese Differenzen als Basis zur Berechnung des Mittelwerttests bei verbundenen Stichproben. Dabei wird die Hypothese überprüft, ob global (für die Gesamtgruppe) eine Veränderung stattgefunden hat. Neben dieser skizzierten Verwendung des Differenzscores zur Analyse von Globalveränderung kann der Differenzwert auch zur Analyse bzw. Prognose *individueller* Veränderung herangezogen werden. Unter welchen Bedingungen erlaubt der Differenzwert dann optimale Schlüsse?

Die Konstruktion eines Index kann aus mehreren Gründen erfolgen. So kann man die Variablenzahl reduzieren wollen oder neue Aussagen beabsichtigen, die mit den Ausgangsvariablen nur schwer möglich wären. Diese zweite Absicht ist aber nur dann realisierbar, wenn eine relativ starke Unabhängigkeit zwischen den Ausgangsvariablen und dem Index existiert. Wie groß diese Unabhängigkeit ist, soll mit Hilfe der Korrelation zwischen Anfangswert und Differenzwert untersucht werden. Analoge Ergebnisse kann man für die Korrelation zwischen Differenzwert und Endwert erreichen.

#### 4.1 Korrelation zwischen Anfangswert und Differenzwert

Falls die Varianz der Anfangswerte:  $V(Y_1)$  größer oder gleich der Varianz der Endwerte  $V(Y_2)$  ist, ist die Korrelation zwischen Anfangswert  $Y_1$  und Differenzwert  $D$  negativ:

$$\varrho(Y_1, D) \leq 0 \text{ bei } V(Y_1) \geq V(Y_2)$$

Denn die Korrelation zwischen  $Y_1$  und  $D = Y_2 - Y_1$  ist:

$$\varrho(Y_1, D) = \varrho(Y_1, Y_2 - Y_1) = \frac{\text{Cov}(Y_1, D)}{\sqrt{V(Y_1) \cdot V(D)}}$$

$$V(D) = V(Y_2 - Y_1) = V(Y_1) + V(Y_2) - 2 \text{Cov}(Y_1, Y_2)$$

$$\begin{aligned} \text{Cov}(Y_1, D) &= \text{Cov}(Y_1, Y_2 - Y_1) = \text{Cov}(Y_1, Y_2) - \text{Cov}(Y_1, Y_1) \\ &= \text{Cov}(Y_1, Y_2) - V(Y_1) \end{aligned}$$

$$\varrho(Y_1, D) = \frac{\text{Cov}(Y_1, Y_2) - V(Y_1)}{\sqrt{V(Y_1)} \sqrt{V(Y_1) + V(Y_2) - 2 \text{Cov}(Y_1, Y_2)}}$$



Im Nenner erhält man die Normierung der Cov ( $Y_1, D$ ). Zur Abschätzung, wann der Korrelationskoeffizient negativ ist, muß der Zähler betrachtet werden.

Allgemein gilt:  $\sqrt{V(Y_1) \cdot V(Y_2)} \geq \text{Cov}(Y_1, Y_2)$

Laut Voraussetzung ist:  $V(Y_1) \geq V(Y_2)$

$$\begin{aligned} V(Y_1) &= \sqrt{V(Y_1) V(Y_1)} \geq \sqrt{V(Y_1) V(Y_2)} \geq \text{Cov}(Y_1, Y_2) \\ 0 &\geq -V(Y_1) + \text{Cov}(Y_1, Y_2) = \text{Cov}(Y_1, D) \\ \rho(Y_1, D) &\leq 0 \text{ falls } V(Y_1) = V(Y_2) \\ &\quad \text{oder } V(Y_1) > V(Y_2) \end{aligned}$$

Falls die Varianz  $V(Y_1) < V(Y_2)$  ist, kann der Korrelationskoeffizient durchaus positiv sein; daher ist die Tendenz zu einer negativen Korrelation (in der Literatur bekannt unter dem Stichwort: „Regression zum Mittelwert“) zwischen Anfangswert und Differenzwert entscheidend bestimmt durch die Varianzen der Werte zu beiden Zeitpunkten. Coleman (1968) charakterisiert diese Tendenz als „negativen Feedback“, der ein System stabil erhält. Bei positiver Korrelation müßte bei Fortschreiten des Prozesses wegen der permanent zunehmenden Varianz ein explosiver Prozeß vorliegen: Je größer die individuelle positive Abweichung vom Anfangsmittelwert ist, desto größer ist die individuelle Zunahme. Entsprechendes gilt spiegelbildlich für individuelle negative Abweichungen,

Als Konsequenz für eine Längsschnittstudie ergibt sich, daß Maßnahmen verhindert werden sollten, die zu einer „künstlichen“ Varianzstabilisierung führen könnten. So sollte man Meßinstrumente, die nur einen beschränkten Ausschnitt eines Bereichs messen (Ceiling effect) oder eine Standardisierung der Variablen für die Zeitpunkte (Normierung von Instrumenten) vermeiden. Zumindest müssen bei der Interpretation eines „negativen Feedbacks“ solche varianzstabilisierenden Maßnahmen mit berücksichtigt werden. Im Allgemeinen kann man von der Abhängigkeit vom Anfangs- und Veränderungswert ausgehen:

## 4.2 Schätzung individueller Veränderungswerte

Ist man an der Prognose individueller Veränderungswerte bzw. der Prognose des jeweiligen Endwertes interessiert, muß man untersuchen, in welchem Ausmaß die Anfangswerte mit berücksichtigt werden sollen.

Für eine solche Prognose könnte man folgende drei Verfahren vorschlagen:

1. a) Berechnung eines globalen Erwartungswertes:  $E(Y_2 - Y_1) = E(D)$
- b) Endwert (geschätzt):  $\hat{y}_{v2} = y_{v1} + E(D)$ ;  $v$  = Personenindex

---

\* Z.B.: Bereiter, 1963.

2. a) Berechnung eines bedingten Erwartungswertes: Regression 1. Art:  
 $E(Y_2 | Y_1 = y_{v1})$  für alle verschiedenen  $y_{v1}$   
 b) Endwert (geschätzt):  $\hat{y}_{v2} = E(Y_2 | Y_1 = y_{v1})$
3. a) Unter der zusätzlichen Annahme, daß der bedingte Erwartungswert von  $Y_2$  eine lineare Funktion der Anfangswerte ist:  
 $E(Y_2 | Y_1 = y_{v1}) = a + by_{v1}$ , kann die Steigung  $b$  und der Abschnitt auf der  $y_2$ -Achse berechnet werden  
 (Regression 2. Art; s.a. Fisz, 1973)  
 b) Endwert (geschätzt):  $\hat{y}_{v2} = a + by_{v1}$

Beispiel: Erholungskur

		Y <sub>1</sub>				
		40	50	60	70	
Gewicht nach Kur: Y <sub>2</sub>	80	0	0	.25	.75	.25
	70	0	.25	.50	.25	.25
	60	.25	.50	.25	0	.25
	50	.75	.25	0	0	.25
		.25	.25	.25	.25	

pro Spalte in der Tabelle: bedingte Wahrscheinlichkeiten bzw. bedingte relative Häufigkeiten

$$E(D) = E(Y_2 - Y_1) = E(Y_2) - E(Y_1) = 65 - 55 = 10$$

$$E(Y_2 | 40) = 60 \cdot .25 + 50 \cdot .75 = 52.5$$

$$E(Y_2 | 50) = 60 \quad E(Y_2 | 60) = 70 \quad E(Y_2 | 70) = 77.5$$

$$a = E(Y_2) - bE(Y_1) = 65 - .85 \cdot 55 = 18.25$$

$$b = \frac{\text{cov}(Y_1, Y_2)}{\sqrt{V(Y_1)} \sqrt{V(Y_2)}} = \frac{106.25}{\sqrt{125} \sqrt{125}};$$

Prognosen für  $y_2$ -Werte bei den 3 Verfahren

bei Y <sub>1</sub>	$\hat{Y}_2$ (Verf. 1)	$\hat{Y}_2$ (Verf. 2)	$\hat{Y}_2$ (Verf. 3)
40	40 + 10 = 50	52.5	18.25 + .85 * 40 = 52.25
50	50 + 10 = 60	60	18.25 + .85 * 50 = 60.75
60	60 + 10 = 70	70	18.25 + .85 * 60 = 69.25
70	70 + 10 = 80	77.5	18.25 + .85 * 70 = 77.75

In diesem Beispiel sind die Varianzen zu beiden Zeitpunkten gleich. Der Anfangsmittelwert ist 55. Methoden 2 und 3 zeigen eine stärkere Regression zum Mittelwert als Methode 1, d.h. die Schätzungen für den 2. Zeitpunkt liegen bei Methode 2 und 3 näher beim Mittelwert als bei Methode 1.

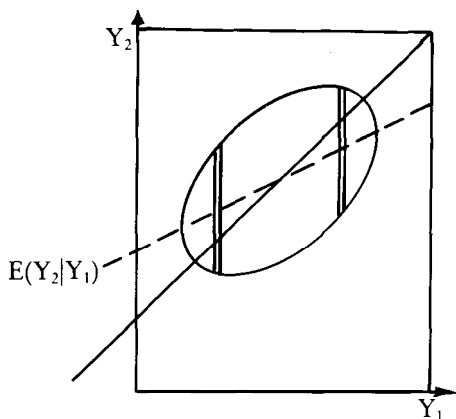


Fig. 4.1: Bivariate Normalvert. mit Hauptachse ——— und Regression  $E(Y_2 | Y_1)$

Als individuelle Abweichungen ergeben sich für die 3 Methoden:

- 1:  $\text{Abw.} = y_{v2} - \hat{y}_{v2} = y_{v2} - y_{v1} - E(D)$  (= Residuen bei Regression  
Konst. ohne Anfangswert)
- 2:  $\text{Abw.} = y_{v2} - \hat{y}_{v2} = y_{v2} - E(Y_2 | y_{v1})$  (= Residuen bei Regression  
1. Art mit Anfangswert)
- 3:  $\text{Abw.} = y_{v2} - \hat{y}_{v2} = y_{v2} - (a + by_{v1})$  (= Residuen in Regression  
 $= y_{v2} - by_{v1} - a$  2. Art mit Anfangswert)  
Konst.

Unter der Bezeichnung „Überkorrektur - Unterkorrektur - Dilemma“ hat Bereiter (1963) das Problem behandelt, wie bei den Differenzwerten die Anfangswerte berücksichtigt und nach den Anfangswerten korrigierte Veränderungswerte erzeugt werden sollen.

In Methode 1 wird durch Subtraktion nach den Anfangswerten korrigiert. Diese Korrektur berücksichtigt aber überhaupt nicht die Art der bivariaten Verteilung zwischen Anfangswert und Endwert. An Verteilungsinformation wird der Mittelwert berücksichtigt, der allerdings nur Informationen über die *Randverteilungen* enthält.

Bei den Methoden 2 und 3 wird die Art der bivariaten Verteilung berücksichtigt. Die Abweichungen sind dabei optimal (im Sinne der Fehlerminimierung) korrigiert.

Bei individuellen Prognosen sollte man entsprechend Cronbach & Furby (1970) nach Möglichkeit alle Informationen berücksichtigen, die man zur Verfügung hat. Darunter fällt auch der Anfangswert. Zur Berücksichtigung der

Information über Meßfehler wurden spezielle Formeln entwickelt. Eine Übersicht findet sich bei Petermann (1978).

Bei *globalen* Fragestellungen („Gibt es in einer gegebenen Population eine Wirkung der Kur?“) ist eine Korrektur nach den Anfangswerten nicht erforderlich. Eine solche Fragestellung sei „*unbedingt*“ genannt. Fragestellungen, bei denen die unterschiedlichen Anfangswerte mitberücksichtigt (etwa: „Welche Wirkung ist für eine bestimmte Person zu erwarten?“) und irgendwie konstant gehalten werden sollen, heißen „*bedingte*“ Fragestellungen (s. Bock, 1975).

### 4.3 Der Differenz- bzw. Endwert in der Regressionsanalyse

Wir wollen jetzt den Einfluß von Drittvariablen auf die Veränderung D untersuchen.

*Differenz- bzw. Endwert als abhängige Variable:*

Für die Form der Beziehung zwischen Differenzwert und einer Drittvariablen wird ein lineares Modell gewählt, wobei der Anfangswert auf der Seite der unabhängigen Variablen auftaucht („bedingte“ Fragestellung).

(4.1)

$$\begin{array}{ccccccc}
 \begin{array}{|c|} \hline D_1 \\ D_2 \\ \dots \\ D_N \\ \hline \end{array} & = & \begin{array}{|c|} \hline Y_{12} - Y_{11} \\ Y_{22} - Y_{21} \\ \dots \\ Y_{N2} - Y_{N1} \\ \hline \end{array} & = & \begin{array}{|c|} \hline Y_{11} \\ Y_{21} \\ \dots \\ Y_{N1} \\ \hline \end{array} & \begin{array}{|c|} \hline b_1 \\ \hline \end{array} & + & \begin{array}{|c|} \hline X_{12} \dots X_{1m} \\ X_{22} \dots X_{2m} \\ \dots \\ X_{N2} \dots X_{Nm} \\ \hline \end{array} & \cdot & \begin{array}{|c|} \hline b_2 \\ b_3 \\ \dots \\ b_m \\ \hline \end{array} & + & \begin{array}{|c|} \hline \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \\ \hline \end{array} \\
 d & = & y_2 - y_1 & = & y_1 & b_1 & + & X & & b & + & \varepsilon
 \end{array}$$

Anfangswert

unabhängige Variable (incl. Scheinvariable mit Werten 1 für Regressionskonstante)

Durch Umformung der Gleichung (4.1) kann man erreichen, daß nur noch der Endwert auf der linken Seite steht:

$$(4.2) \quad y_2 = y_1 b_1 - y_1 + Xb + \varepsilon = y_1(b_1 - 1) + Xb + \varepsilon$$

Man sieht, daß die Regressionskoeffizienten der Drittvariablen unverändert bleiben. Eine einfache Umformung erfährt nur der Regressionskoeffizient des Anfangswerts. Dies trifft auch für die Schätzungen zu. Da auch die Varianzen der Schätzer in beiden Fällen gleich sind, ist es für die Praxis irrelevant, ob man den Endwert oder den Differenzenwert als abhängige Variable wählt.

Diese Überlegungen sind auch auf Gleichungssysteme (s. Roskam, 1979) und auf Regressionen mit fehlerbehafteten Variablen (Werts & Linn, 1970) übertragbar. Jedoch gilt diese Invarianz nur für den Fall der bedingten Fragestellung. Anderweitig können sich alle Schätzungen für die Koeffizienten  $b_i$  ( $i=2, \dots, m$ ) ändern.

Bei unbedingter Fragestellung gilt:

$$(4.3) \quad b_D = b_E - b_A \quad \text{wobei:}$$

$b_D$  Regressionskoeffizienten  
bei Differenz als abhängige Variable  
 $b_A$  bei Anfangswert als abhängige Variable  
 $b_E$  bei Endwert als abhängige Variable

Die unabhängigen Variablen müssen identisch sein.

### *Der Differenzwert als unabhängige Variable*

Es sei  $Z$  eine abhängige Variable (z.B. subjektive Befindlichkeit nach einer Kur) und die unabhängigen Variablen bestünden aus dem Anfangswert  $Y_1$ , dem Differenzwert  $D$  und anderen Drittvariablen  $X_i$  ( $i=2, \dots, m$ ).

Die multiple Regression sei in Matrixform gegeben:

$$(4.4) \quad z = y_1 b_0 + d b_1 + Xb + \varepsilon = y_1 b_0 + (y_2 - y_1) b_1 + Xb + \varepsilon$$

Durch Umformung von (4.4) kann man erreichen, daß nicht mehr die Differenz  $(y_2 - y_1)$  sondern nur noch der Anfang- und Endwert in der Gleichung vorkommen:

$$(4.5) \quad z = y_1 b_0 - y_1 b_1 + y_2 b_1 + Xb + \varepsilon = y_1 (b_0 - b_1) + y_2 b_1 + Xb + \varepsilon$$

Man sieht wiederum, daß die Regressionskoeffizienten der Drittvariablen  $X$  invariant gegenüber der Wahl von Differenz oder Endwert auf der Prädiktorseite sind. Dasselbe gilt für die Schätzer  $b$  und deren Varianzen.

### *Partielle Korrelation zwischen Differenzwert und Drittvariablen bei bedingter Fragestellung*

Im Spezialfall einer einzigen Drittvariablen  $X$  gilt:

$$(4.6) \quad Q_{DX.Y_1} = Q_{Y_2X.Y_1} \quad \text{wobei: } Q_{DX.Y_1} = \text{partielle Korrelation}$$

zw. Differenzwert und  
Drittvariablen unter  
Konstanthaltung des  
Anfangswertes  $Y_1$

Dieses Ergebnis ist nicht weiter überraschend, da die Regressionskoeffizienten in den entsprechenden Gleichungen (4.2) und (4.4) invariant gegenüber der Wahl von Differenz- oder Endwert sind. Diese Invarianz überträgt sich dabei auf die partiellen Korrelationskoeffizienten (s. Yule & Kendall, 1964).

### *Vergleich von bedingter und unbedingter Fragestellung*

Der Vergleich ist bekannt geworden unter dem Begriff des „Lord’schen Paradoxon“. Dabei wird einer unbedingten Analyse von Differenzwerten eine Kovarianzanalyse der Endwerte (mit den Anfangswerten als Kovariabler) gegenübergestellt.

Lord bezieht sich in seinem Beispiel auf die Gewichtsveränderung von Jungen und Mädchen im Zeitraum von September bis Juni (Bock, 1975, S. 490f.):

#### *Lord’s paradox:*

Suppose a large university obtains measurements, at the beginning and end of the school year, of the weight of each student who takes his meals in the university dining halls. When the resulting data are classified by sex of student, their scatter plot takes the form shown schematically in Figur 4.2. The 45 line represents equality of weights in September and June. The ellipses of concentration represent the presumably bivariate normal distribution of weight on the two occasions.

Suppose two statisticians analyse these data for differences in weight gain of men versus women. The *first statistician* analyses simple gain scores and concludes that “as far as these data are concerned, there is no evidence of any interesting effect of the school diet (or of anything else) on student weight; in particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change.”

The *second statistician*, on the other hand, decides to do an analysis of covariance.

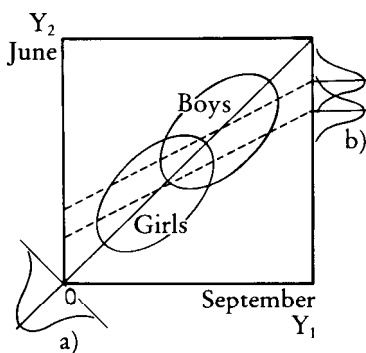


Fig. 4.2: Verteilung der Zunahme bei a) unbedingter, b) bedingter Analyse

After some necessary preliminaries, he determines that the slope of the regression line of final weight on initial weight is essentially the same for the two sexes. This is fortunate since it makes possible a fruitful comparison of the intercepts of the regression lines. . . He finds that the difference between the intercepts is statistically highly significant. The second statistician concludes . . . that the [men] showed significantly more gain in weight than the [women] when proper allowance is made for differences in initial weight between the two sexes.

In beiden Arten der Analyse werden als unabhängige Variable eine Konstante und das Geschlecht eingeführt. Bei der Kovarianzanalyse wird zusätzlich der Anfangswert als unabhängige Variable berücksichtigt. Als abhängige Variable werden bei der Analyse der Differenzenwerte die Differenzen, bei der Kovarianzanalyse die Endwerte verwendet. Die Fragestellung lautet: „Sind noch Geschlechtsunterschiede festzustellen?“

Der Lord'sche Differenzenansatz kann auch in die Form des allgemeinen linearen Modells gebracht werden.

$$(4.7) \quad \begin{array}{c} \delta \\ \vdots \\ \delta \\ \vdots \\ \delta \\ \vdots \\ \delta \end{array} \left\{ \begin{array}{c} D_1 \\ \dots \\ D_1 \\ D_{I+1} \\ \dots \\ D_N \end{array} \right\} = \begin{array}{cc} \begin{array}{cc} 1 & 0 \\ \dots & \dots \\ 1 & 0 \\ 0 & 1 \\ \dots & \dots \\ 0 & 1 \end{array} & \begin{array}{c} b_M \\ b_W \end{array} \end{array} + \begin{array}{c} \epsilon_1 \\ \dots \\ \epsilon_I \\ \epsilon_{I+1} \\ \dots \\ \epsilon_N \end{array}$$

Die Fragestellung kann durch die Hypothese:

$$H_0 : b_M = b_W (\equiv b_M - b_W = 0) \quad \text{oder auch: } \frac{b_M - b_W}{2} = 0$$

repräsentiert werden. Man kann die Regressionsgleichung für diese Fragestellung reparameterisieren, da nur der Effektunterschied bezüglich des Geschlechts interessiert:

$$(4.8) \quad \begin{array}{c} \delta \\ \vdots \\ \delta \\ \vdots \\ \delta \\ \vdots \\ \delta \end{array} \left\{ \begin{array}{c} D_1 \\ \dots \\ D_1 \\ D_{I+1} \\ \dots \\ D_N \end{array} \right\} = \begin{array}{cc} \begin{array}{cc} 1 & 1 \\ \dots & \dots \\ 1 & 1 \\ 1 & -1 \\ \dots & \dots \\ 1 & -1 \end{array} & \begin{array}{c} a_G \\ a_U \end{array} \end{array} + \begin{array}{c} \epsilon_1 \\ \dots \\ \epsilon_I \\ \epsilon_{I+1} \\ \dots \\ \epsilon_N \end{array}$$

Dabei bedeutet dann

$$a_G = \frac{b_M + b_W}{2}$$

$$a_U = \frac{b_M - b_W}{2}$$

(s. Bock, 1975)

Der interessierende Effektunterschied ist dann durch den Parameter  $a_U$  repräsentiert.

Durch den Test der Hypothese  $H_0 : a_U = \frac{1}{2} (b_M - b_W) = 0$  kann die unbedingte Fragestellung beantwortet werden.

Der angeführte Kovarianzansatz berücksichtigt den Anfangswert als Unabhängige:

$$(4.9) \quad \begin{matrix} \sigma \\ \varphi \end{matrix} \begin{Bmatrix} Y_{12} \\ \dots \\ Y_{12} \\ Y_{1+1,2} \\ \dots \\ Y_{N2} \end{Bmatrix} = \begin{bmatrix} 1 & 1 \\ \dots & \dots \\ 1 & 1 \\ 1 & -1 \\ \dots & \dots \\ 1 & -1 \end{bmatrix} + \begin{bmatrix} c_G \\ c_U \end{bmatrix} + \begin{Bmatrix} Y_{11} \\ \dots \\ Y_{11} \\ Y_{1+1,1} \\ \dots \\ Y_{N1} \end{Bmatrix} + \begin{bmatrix} c_1 \\ \dots \\ c_1 \\ \dots \\ \dots \end{bmatrix} + \begin{Bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_1 \\ \varepsilon_{1+1} \\ \dots \\ \varepsilon_N \end{Bmatrix}$$

Die Untersuchung der Fragestellung kann dabei wieder durch die Hypothese  $H_0 : c_U = 0$  geprüft werden.

Wie schon oben erwähnt, verändern sich weder die Parameterwerte  $c_G$ ,  $c_U$  noch deren Varianzen, wenn man als abhängige Variable den Endwert  $Y_2$  durch die Differenz  $D$  ersetzt (nur bei  $c_1$  ergäben sich Unterschiede). Aus diesem Grund liegt der Unterschied in den beiden Analyseformen nicht in der Verwendung von Differenz- oder Endwert als abhängiger Variablen, sondern nur in Aufnahme oder Nichtaufnahme des Anfangswertes als zusätzlicher Kovariaten begründet. Entscheidend ist also die Wahl einer bedingten oder unbedingten Fragestellung.

### Regressionskoeffizienten bei bedingter und unbedingter Fragestellung

Die Unterschiede zwischen den Regressionskoeffizienten bei bedingter und unbedingter Fragestellung ergeben sich aus der Betrachtung, eine weitere unabhängige Variable in einer Regressionsanalyse wegzulassen bzw. hinzuzunehmen.

Das Regressionsmodell für die bedingte Fragestellung ist:

$$(4.10) \quad \begin{matrix} 1 \\ \vdots \\ \vdots \\ \vdots \\ N \end{matrix} \begin{bmatrix} y_2 \end{bmatrix} = \begin{matrix} 1 & 2 \dots m \\ \begin{bmatrix} y_1 & X \end{bmatrix} \end{matrix} \cdot \begin{bmatrix} b_1 \\ b \end{bmatrix} + \begin{bmatrix} \varepsilon \end{bmatrix} \quad \begin{bmatrix} b_1 \\ b \end{bmatrix} = \text{Vektor mit Regressionskoeffizienten}$$



und das Regressionsmodell für die unbedingte Fragestellung (ohne Anfangswerte  $Y_1$ )

(4.11)

1

⋮

N

$y_2$

=

$X$

·

$c$

+

$\varepsilon$

$c$

 = Vektor mit  
Regressions-  
koeffizienten

Dann gilt die einfache Umrechnungsformel:

(4.12)  $\hat{b} = \hat{c} - \hat{s}\hat{b}_1$       bzw.  $\hat{c} = \hat{b} + \hat{s}\hat{b}_1$

Dabei ist  $\hat{s}$  der geschätzte Regressionsvektor der Hilfsgleichung:

(4.13)  $y_1 = Xs + \varepsilon''$

Man sieht, daß nur dann  $\hat{c} = \hat{b}$  sein kann, wenn einerseits der Regressionskoeffizient  $b_1$  im Rahmen der bedingten Analyse (4.10) Null ist, oder andererseits  $s$  in (4.13) dem Nullvektor entspricht. Erweiterungen auf mehrere unterdrückte Variable sind ebenfalls möglich (Nagl, 1983). Die gleichen Überlegungen gelten auch für die Erwartungswerte (s.a. „Spezifikationsanalyse“ von Theil, 1971).

### 4.4 Kovarianz- bzw. Regressionsmodell bei zeitbezogenen Daten

Mit Hilfe der Regressionsanalyse<sup>\*</sup> kann man allgemein die direkten Effekte der „unabhängigen“ Variable auf die abhängige isolieren: jeder Regressionskoeffizient stellt den Effekt der unabhängigen auf die abhängige Variable unter Konstanthaltung der übrigen Regressoren dar. Die „unabhängigen“ Variablen selber dürfen korreliert sein (Fig. 4.3) (S. 339).

Voraussetzung für eine richtige Konstanthaltung ist aber, daß keine Fehlspezifikation des Modells oder Verletzungen der Modellannahmen vorliegen.

- Hierbei sind unter anderem\*\* folgende Fehlermöglichkeiten zu beachten:
- Es könnte sein, daß bei einer Kovarianzanalyse für verschiedene Gruppen, die z.B. durch die qualitativen Ausprägungen der Faktoren gebildet wurden,

---

<sup>\*</sup> Da das Kovarianzmodell bei fixen Faktoren als Spezialfall des Regressionsmodells formulierbar ist (z.B. Searle, 1971), gelten die folgenden Bemerkungen über die Regressionsanalyse auch für die Kovarianzanalyse.

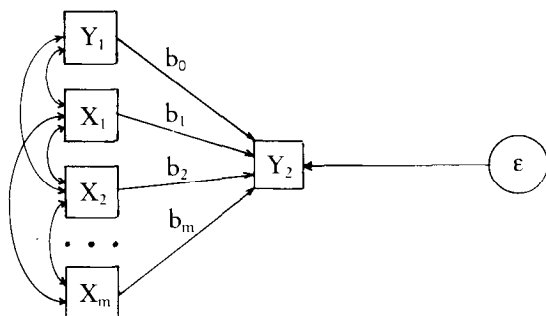
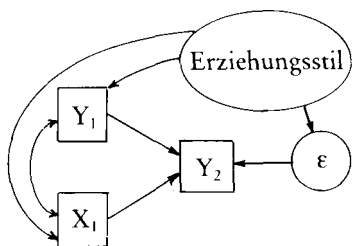


Fig. 4.3: Graph für multiple Regression

den, unterschiedliche Steigungen zutreffen (hierfür gibt es den Test für die Homogenität der Regressionen s. Bock, 1975). In diesem Fall müßte eine Intra-Class-Regression durchgeführt werden (s. Searle, 1971, S. 355), bei der die Steigungen für die Gruppen unterschiedlich sein könnten.

- Wichtige Regressoren (auch Potenzen von oder Interaktionen zwischen Regressoren) könnten im Modell fälschlicherweise nicht berücksichtigt worden sein.
- Der dritte mögliche Fehler ist wohl der gravierendste. Es könnte sein, daß die Störgröße  $\epsilon$  mit einem Regressor (= unabhängige Variable) kovariert. In diesem Fall ist eine wesentliche Grundvoraussetzung für das Schätzen der Regressionskoeffizienten verletzt (s. Goldberger, 1964).



Erziehungsstil wird fälschlich aus der Analyse ausgelassen

$Y_1$ : Angst mit 10 J.

$Y_2$ : Angst mit 15 J.

$X_1$ : Selbstsicherheits-training : 1

Kontrollgruppe : 0

Fig. 4.4: „wahres“ Kausalmodell (in die statistische Analyse werden fälschlich nur die Variablen  $X_1, Y_1$  und  $Y_2$  aufgenommen)

\*\*\* Weitere Fehlspezifikationen liegen vor, wenn etwa die Annahme der Homogenität der Varianzen der verschiedenen  $\epsilon$  nicht gegeben ist (siehe dazu: Goldberger, 1964), oder die Effekte mit den Regressoren korrelieren (siehe dazu: Werts & Linn, 1971).

Da  $\varepsilon$  nur ein Störglied ist, das selbst wiederum eine Menge von Einflüssen repräsentiert (s. Theil, 1971), ist es möglich, daß (a) eine Komponente von  $\varepsilon$  eine „unabhängige“ Variable (z.B.  $Y_1, X_1$ ) beeinflusst (s. Figur 4.4), (b) eine Komponente von  $\varepsilon$  von einer externen Größe (z.B. Erziehungsstil) beeinflusst wird, die auch auf eine unabhängige Variable (z.B.  $X_1, Y_1$ ) „wirkt“; (c) eine unabhängige Variable (z.B.  $X_1, Y_1$ ) eine Komponente von  $\varepsilon$  beeinflusst; (d) Meßfehler in den „unabhängigen“ Variablen (z.B.  $Y_1, X_1$ ) zu einer Komponente von  $\varepsilon$  werden (s.a. u.).

Alle vier Möglichkeiten (a)-(d) führen zur Kovariation zwischen  $\varepsilon$  und den unabhängigen Variablen und führen zu inkonsistenten Schätzungen. Mögliche Abhilfen wären:

### 1. Randomisierung:

Es wird per Zufall entschieden, wer im Selbstsicherheitstraining aufgenommen wird. Dadurch wird es sehr wahrscheinlich, daß die Variable „Training/Nichttraining“ mit keiner anderen Variablen zum Zeitpunkt der Randomisierung korreliert. Gleiches gilt auch für die Korrelation mit Komponenten von  $\varepsilon$ . Darüber hinaus braucht man im allgemeinen auch keine Kovariate (hier: Angst mit 10 Jahren) mehr, nach der bereinigt werden soll. Mit großer Wahrscheinlichkeit korreliert die Trainingsvariable auch nicht mehr mit der Kovariaten. Eine ausführliche Behandlung dieses Gedankens findet sich bei Miller (1971). Sie zeigt allerdings auch, wie neue Schätzprobleme durch die Randomisierung auftauchen können.

Aber auch aus ethischen und therapeutischen Gründen kann das aus statistischen Überlegungen wohl wünschenswerte Randomisieren problematisch sein.

### 2. Einführung weiterer unabhängiger Variabler in die Regression:

Hierdurch kann man unter Umständen ebenfalls erreichen, daß die Kovarianz verschwindet.

### 3. Berücksichtigung von Meßfehlern:

Im Rahmen der Regressionsanalyse sollte berücksichtigt werden, daß nicht alle Variablen exakt gemessen werden können. Die gemessenen manifesten Variablen ( $Y_i$ ) setzen sich dann zusammen aus einem wahren Wert ( $\eta_i$ ) und einem Meßfehler ( $\varepsilon_i$ ). Man ist oft interessiert an der Beziehung zwischen den „wahren“ Werten (s. Figur 4.5).

Durch Einsetzung von

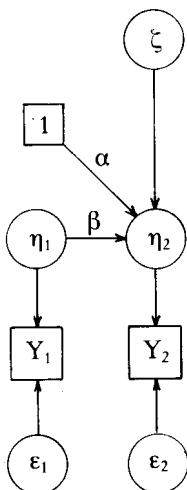
$$\eta_1 = Y_1 - \varepsilon_1, \eta_2 = Y_2 - \varepsilon_2 \quad \text{in} \quad \eta_2 = \alpha + \beta \eta_1 + \zeta$$

erhält man :

$$Y_2 = \alpha + \beta Y_1 + \underbrace{(\varepsilon_2 - \beta \varepsilon_1)}_{\text{Störglied bei üblicher Regressionsanalyse}} + \zeta$$

Störglied bei üblicher Regressionsanalyse

Graphisch:



algebraisch:

(4.14)	a) $Y_1 = \eta_1 + \varepsilon_1$	Meß- konzeption
	b) $Y_2 = \eta_2 + \varepsilon_2$	
	c) $\eta_2 = \alpha + \beta \eta_1 + \zeta$	

Dabei wird angenommen,  
daß  $\varepsilon_1, \varepsilon_2$  jeweils  
nicht mit  $\eta_1, \eta_2$  korre-  
lieren und  $\text{cov}(\varepsilon_1, \varepsilon_2) = 0$   
(Klassische Annahme) ist.

Zudem wird angenommen, daß auch  
das Störglied ( $\zeta$ ) nicht mit  $\eta_1$ ,  
 $\varepsilon_1$  und  $\varepsilon_2$  korreliert.

Fig. 4.5: Pfadmodell für 2 Zeitpunkte

Man kann zeigen, daß dieses Störglied mit  $Y_1$  (= unabhängige Variable bzw. Pretest) korreliert. Daher sind (wie oben dargelegt) die üblichen Regressionsverfahren nach der Methode der kleinsten Quadrate nicht anwendbar (s.a. Goldstein, 1979). Allerdings sind ML-Schätzungen möglich, wie Kendall & Stuart II (1973) gezeigt haben. Sie beschränken sich aber im Gegensatz zu uns auf den Fall  $\sigma_{\zeta}^2 = 0$ . Die Schätzgleichungen lauten:

(4.15)	a) $E(\eta_1) = \mu \triangleq \bar{Y}_1$ (= Mittelwert)	
	b) $E(\eta_2) = \alpha + \beta \mu \triangleq \alpha + \beta \bar{Y}_1$	$\sigma_{\eta_1}^2$ = Varianz ( $\eta_1$ )
	c) $\sigma_{\eta_1}^2 + \sigma_{\varepsilon_1}^2 \triangleq S_{y_1}^2$	$S_{y_1}^2$ = Varianz von $Y_1$ in der Stichprobe
	d) $\beta^2 \cdot \sigma_{\eta_1}^2 + \sigma_{\zeta}^2 + \sigma_{\varepsilon_2}^2 \triangleq S_{y_2}^2$	
	e) $\beta \cdot \sigma_{\eta_1}^2 \triangleq \widehat{\text{cov}}(Y_1, Y_2) = S_{Y_1 Y_2}$	Analog für $\sigma_{\eta_2}^2$ und $S_{y_2}^2$ $\left\{ \begin{array}{l} \text{Kovarianz von } Y_1 \text{ und} \\ Y_2 \text{ in der Stichprobe} \end{array} \right.$

Insgesamt sind fünf Gleichungen in sieben Unbekannten  $\mu, \alpha, \beta, \sigma_{\eta_1}^2, \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_2}^2, \sigma_{\zeta}^2$  gegeben. Daraus sieht man, daß man ohne Beschränkungen keine eindeutige Lösung für die 7 Unbekannten erhalten kann. Daher sind für die Lösung Zusatzannahmen erforderlich. Die beiden ersten Gleichungen reichen gerade aus,  $\mu$  und  $\alpha$  zu berechnen, falls  $\beta$  gegeben ist. Wir beschränken uns daher auf die Berechnung von  $\beta$ :

*Situation A:* Fehlervarianzen bekannt, Correction of Attenuation.

Angenommen, es seien  $\sigma_{\epsilon 1}^2$  und  $\sigma_{\epsilon 2}^2$  bekannt. Das bedeutet, daß man die Reliabilität der Meßinstrumente kennt. Man kann auf Grund der Gleichungen:

$$\sigma_{\eta 1}^2 = \sigma_{y 1}^2 - \sigma_{\epsilon 1}^2 \text{ und } \beta \cdot \sigma_{\eta 1}^2 = \sigma_{y 1 y 2}$$

den Schätzer  $\hat{\beta}$  für  $\beta$  erhalten:

$$\hat{\beta} = \frac{S_{y 1 y 2}}{S_{y 1}^2 - \sigma_{\epsilon 1}^2} \quad \text{Diesen Schätzer könnte man auch mit Hilfe der „Correction of Attenuation“ erhalten (siehe Isaak (1970)).}$$

Aus der Gleichung  $\sigma_{\xi}^2 = \sigma_{y 2}^2 - \sigma_{\epsilon 2}^2 - \beta^2 \sigma_{\eta 1}^2$  kann man, ohne zu einem Widerspruch zu gelangen,  $\hat{\sigma}_{\xi}^2$  berechnen.

### *Situation B: Strukturelle Relation*

Es wird angenommen, daß  $\sigma_{\xi}^2 = 0$  ist. Damit enthält die Relation in den wahren „latenten“ Variablen kein Störglied  $\xi$ . Die möglicherweise notwendige Konstante wird von  $a$  absorbiert. Kendall & Stuart wählen den Begriff „Strukturelle Relation“. Die Beziehung zwischen den Variablen soll *exakt* sein. Ungenauigkeiten kommen nur durch Meßfehler zustande.

Für Schätzungen für  $\beta$  sind etliche Spezialfälle zu unterscheiden, da ja immer noch 6 Unbekannte bei 5 Gleichungen vorhanden sind. Je nach Annahme, resultiert ein anderer Schätzer (s.a. Kendall & Stuart, 1973<sup>3</sup>; Isaak, 1970). Wird etwa angenommen, daß  $\sigma_{\epsilon 1}^2$  (Meßfehlervarianz zum 1. Zeitpunkt) bekannt ist, dann gilt wie oben

$$\hat{\beta} = \frac{S_{y 1 y 2}}{S_{y 1}^2 - \sigma_{\epsilon 1}^2} \quad \text{falls } \sigma_{\epsilon 1}^2 < S_{y 1}^2$$

Ist dagegen  $\sigma_{\epsilon 2}^2$  bekannt, resultiert als Lösung:

$$\hat{\beta} = \frac{S_{y 2}^2 - \sigma_{\epsilon 2}^2}{S_{y 1 y 2}} \quad \text{falls } \sigma_{\epsilon 2}^2 < S_{y 2}^2$$

Falls aber Annahmen über das Verhältnis der Meßfehlervarianzen  $\sigma_{\epsilon 1}^2 = \lambda \sigma_{\epsilon 2}^2$  gemacht werden (wenn bei einem Längsschnitt angenommen werden darf, daß mit demselben Meßinstrument gemessen wurde:  $\sigma_{\epsilon 1}^2 = \sigma_{\epsilon 2}^2$  und  $\lambda = 1$ ) resultiert eine sehr komplexe Formel für  $\beta$ . Können dagegen  $\sigma_{\epsilon 1}^2$  und  $\sigma_{\epsilon 2}^2$  als bekannt vorausgesetzt werden, ist das Gleichungssystem überbestimmt und es empfiehlt sich ein iteratives Verfahren (siehe Kendall & Stuart, 1973). Die resultierenden Schätzer sind ML-Schätzer und damit konsistent.

In der Praxis lassen sich die Annahmen oft nicht gewinnen bzw. begründen, die zur Schätzung von  $\beta$  und  $a$  notwendig sind. Ein Ausweg bietet sich bei Einführung von Paralleltests (bzw. kongenerischen Tests) (Werts, Jöreskog &

Linn (1972) und Werts & Linn (1972)) oder bei Messungen zu mehr als 2 Zeitpunkten (z.B.: Wiley (1970), Jöreskog (1979)). Werden die Schätzer dann noch mit dem Computerprogramm LISREL iterativ bestimmt, erhält man datenadäquate  $\beta$  und  $\hat{\alpha}$ .

## 4.5 Reliabilität - Stabilität

Bereiter (1963) hat angenommen, daß die Korrelation zwischen Anfangswert ( $Y_1$ ) und Endwert ( $Y_2$ ) ein Maß für die Reliabilität sei. Schreibt man die Korrelation mit Hilfe der oben bereitgestellten Formeln:

$$(4.16) \quad r_{(y_1, y_2)} = \frac{S_{y_1 y_2}}{\sqrt{S_{y_1}^2 \cdot S_{y_2}^2}} \quad (\text{durch Einsetzen der Formeln (4.15)})$$

$$\triangleq \frac{\beta \sigma_{\eta_1}^2}{\sqrt{(\sigma_{\eta_1}^2 + \sigma_{\varepsilon_1}^2) (\beta^2 \sigma_{\eta_1}^2 + \sigma_{\xi}^2 + \sigma_{\varepsilon_2}^2)}}$$

scheint es zweifelhaft, daß (4.16) die Formel für die Reliabilität ist. Falls allerdings:

$\sigma_{\varepsilon_1}^2 = \sigma_{\varepsilon_2}^2$  (eine Annahme, die wohl sinnvoll ist: die Fehlervarianzen des Meßinstrumentes bleiben für die Zeitpunkte gleich)

und

$\beta = 1$  (Regressionskoeffizient = 1)

und

$\sigma_{\xi}^2 = 0$  (keine Störgrößen wirken auf den wahren Wert bei 2. Messung ein)

als Annahmen gelten, kann man die Korrelation als Reliabilität deuten:

Dann wird aus (4.16)

$$\frac{\sigma_{\eta_1}^2}{\sqrt{(\sigma_{\eta_1}^2 + \sigma_{\varepsilon_1}^2) (\sigma_{\eta_1}^2 + \sigma_{\varepsilon_1}^2)}} = \frac{\sigma_{\eta_1}^2}{(\sigma_{\eta_1}^2 + \sigma_{\varepsilon_1}^2)} = \frac{\text{Varianz der „wahren“ Werte}}{\text{Varianz der gemessenen Werte}}$$

(siehe Wiley (1970)). Besonders die Annahmen  $\beta=1$  und  $\sigma_{\xi}^2=0$  sind problematisch. Denn diese Parameter ermöglichen Aussagen über die Stabilität der wahren Variable über die Zeit hinweg. Unter welchen Annahmen eine solche Trennung von Reliabilität und Stabilität (auch bei Verletzungen von klassischen Meßannahmen) möglich ist, haben Wheaton et al. (1977) zusammenfassend dargestellt.

## 5. Wachstumskurven- und Varianzanalyse

Zur Analyse des Trends über mehrere Zeitpunkte hinweg kann die Varianzanalyse verwendet werden (siehe Winer (1971)). Dabei kann man verschiedene Arten der Beschreibung von Trends unterscheiden (siehe Bock (1975)). Der Faktor „Zeit“ selbst wird bei Trendüberlegungen meist als fixer Faktor konzipiert (siehe Gaito I. & Wiley D. E. (1963), Bock (1975)).

Im klassischen Modell werden restriktive Annahmen für die Fehlerstruktur gemacht (siehe Kap. 5.1.1), so daß bei Verletzungen dieser Fehlerstruktur nur noch konservative approximative Tests möglich sind (Greenhouse & Geisser, 1959). Daher hat Bock 1963 (siehe auch Bock (1975)) Modelle vorgeschlagen, die auf multivariaten Konzepten beruhen und die er selbst als *ungewichtete (exakte) multivariate Varianzanalyse* bezeichnet (Bock, 1979). Diese Analyse hat allerdings den Nachteil, daß bei unvollständiger Spezifikation (siehe bei (5.19)) des Trends keine effizienten Schätzer resultieren. Die Wachstumskurvenanalyse (Khatrı (1966), Potthoff & Roy (1964)) liefert auch bei nicht vollständiger Spezifikation des Trends effiziente Schätzer. Bock (1979) bezeichnet sie auch als *„gewichtete multivariate Varianzanalyse“*.

Wir werden diese Analyseformen darstellen, wobei einerseits die Fehlerannahmen des klassischen Modells kritisch untersucht und dann durch allgemeinere Annahmen ersetzt werden.

### 5.1 Der Eingruppenfall

#### 5.1.1 Der „wiederholte Messungen“-Ansatz ( $T \geq 2$ , $G = 1$ , $N > 1$ )

Im Rahmen von Überlegungen zur Analyse von Experimenten ist es üblich, die Messungen zu den verschiedenen Zeitpunkten bei verschiedenen Personen als Realisierung folgenden linearen Modells anzusehen:

$$(5.1) \quad Y_{vt} = \mu + \pi_v + \tau_t + \varepsilon_{vt}$$

Die Meßwerte  $Y_{vt}$  werden als lineare Funktion folgender Effekte konzipiert:

1. Generelles Niveau (oft auch als Konstante, bzw. generelles Mittel bezeichnet).
2. Effekte der Zeitpunkte  
 $\tau_t$  ( $t=1, \dots, T$ ) : konstant über Personen
3. Personenniveau ( $\pi_v$  :  $v=1, \dots, N$ ) : konstant über die Zeitpunkte.

4. Störglied ( $\epsilon_{vt} : t=1,2; v=1,\dots,N$ ) für jede Person und jeden Zeitpunkt. Dieser Parameter kann auch als Interaktion zwischen Zeit und Personen interpretiert werden (siehe: Winer, 1971).

Dieses Modell ist ein Spezialfall des in der Varianzanalyse standardmäßig verwendeten Modellansatzes.

Zur Diskussion der Parameter läßt sich folgendes anmerken:

- Zu 4.) Die Störgröße  $\epsilon_{vt}$  wird von vornherein üblicherweise als Zufallsgröße konzipiert, die einen Erwartungswert von 0 und eine Varianz von  $\sigma_\epsilon^2$  für alle  $\epsilon_{vt}$  hat:

$$(5.2) \quad \epsilon_{vt} \sim (0, \sigma_\epsilon^2).$$

Ebenfalls wird angenommen, daß die  $\epsilon_{vt}$ -Größe unkorreliert mit jeder anderen  $\epsilon_{v't'}$  ist.

- Zu 3.) Der Personeneffekt wird im Rahmen des „gemischten Modells“ als Zufallsgröße konzipiert. Dabei interessiert nicht der einzelne Effekt, sondern die Varianz dieser Zufallsgröße. Es wird angenommen, daß die Varianz für alle N Effekte gleich ist:

$$(5.3) \quad \text{Var}(\pi_v) = \sigma_\pi^2 \quad (v=1,\dots,N)$$

Weiter wird angenommen, daß die Kovarianz zwischen verschiedenen Effekten 0 ist:

$$\text{Cov}(\pi_v, \pi_{v'}) = 0 \quad (v \neq v'; v, v' = 1, \dots, N)$$

und auch:

$$(5.4) \quad \text{Cov}(\pi_v, \epsilon_{vt}) = 0$$

Der Erwartungswert von jedem  $\pi_v$ :

$$(5.5) \quad E(\pi_v) = 0$$

- Zu 2.) Bei einer Analyse von Trends sind die Effekte der Zeitpunkte von zentralem Interesse, manchmal werden allerdings eher andere Arten der Darstellungen gesucht, als jedem Zeitpunkt selbst einen Effekt bestimmter Größe zuzuordnen (siehe auch Reparametrisierungen, Kap. 5.1.2). Bei Trendanalysen nimmt man meist den Zeitfaktor als fix an.

Für die möglichen  $Y_{vt}$ -Meßwerte bei gegebener Anzahl von Personen (=N) und Zeitpunkten (=T) ergibt sich explizit die Vektor- bzw. Matrixform (ausgerollte Form):



Zeit-  
effekte

Personen-  
effekte

Störeffekte  
(= Fehler)

1. Person  
  
  
  
  
  
  
  
  
  
N. Person

$\begin{pmatrix} Y_{11} \\ \dots \\ Y_{1T} \\ \dots \\ Y_{N1} \\ \dots \\ Y_{NT} \end{pmatrix}$

=

$\begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}_T$

\cdot \mu +

$\begin{pmatrix} \tau_1 \\ \dots \\ \tau_T \\ \dots \\ \tau_1 \\ \dots \\ \tau_T \end{pmatrix}$

+ \dots +

$\begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix} \cdot \pi_N$

+

$\begin{pmatrix} \varepsilon_{11} \\ \dots \\ \varepsilon_{1T} \\ \dots \\ \varepsilon_{N1} \\ \dots \\ \varepsilon_{NT} \end{pmatrix}$

Für die v-Person:

$\begin{pmatrix} Y_{v1} \\ Y_{v2} \\ \dots \\ Y_{vT} \end{pmatrix}$

=

$\begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$

\cdot \mu +

$\begin{pmatrix} \tau_1 \\ \tau_2 \\ \dots \\ \tau_T \end{pmatrix}$

+

$\begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix} \cdot \pi_v$

+

$\begin{pmatrix} \varepsilon_{v1} \\ \varepsilon_{v2} \\ \dots \\ \varepsilon_{vT} \end{pmatrix}$

(5.6)

$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$

=

$\begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \dots \\ \tau_T \end{pmatrix}$

+

$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$

=

$\begin{pmatrix} \pi_v \\ \varepsilon_{v1} \\ \varepsilon_{v2} \\ \dots \\ \varepsilon_{vT} \end{pmatrix}$

$y_v =$

$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$

\cdot

$H$

+

$\begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \dots \\ \tau_T \end{pmatrix}$

+

$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$

\cdot

$H$

\cdot

$\begin{pmatrix} \pi \\ \varepsilon \end{pmatrix}$

Es ist auch oft üblich die Darstellung (5.6) als die „ausgerollte Form folgender Matrixschreibweise zu bezeichnen (s. Timm, 1975; McDonald & Swaminathan, 1973) :

(5.7a)

$N \cdot \begin{pmatrix} Y_{11} \dots Y_{1T} \\ Y_{21} \dots Y_{2T} \\ \dots \\ Y_{v1} \dots Y_{vT} \\ \dots \\ Y_{N1} \dots Y_{NT} \end{pmatrix}$

=

$\begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \\ \dots \\ 1 \end{pmatrix}$

\cdot

$\begin{pmatrix} \mu + \tau_1, \dots, \mu + \tau_T \end{pmatrix}$

+

$\begin{pmatrix} \pi_1 \\ \pi_2 \\ \dots \\ \pi_v \\ \dots \\ \pi_N \end{pmatrix}$

+

$\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}$

+

$\begin{pmatrix} \varepsilon_{11} \dots \varepsilon_{1T} \\ \varepsilon_{21} \dots \varepsilon_{2T} \\ \dots \\ \varepsilon_{v1} \dots \varepsilon_{vT} \\ \dots \\ \varepsilon_{N1} \dots \varepsilon_{NT} \end{pmatrix}$

(5.7b)

$\begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \\ \dots \\ 1 \end{pmatrix}$

=

$\begin{pmatrix} \mu, \tau_1, \dots, \tau_T \end{pmatrix}$

\cdot

$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

+

$\begin{pmatrix} \pi_1 & \varepsilon_{11} & \dots & \varepsilon_{1T} \\ \pi_2 & \varepsilon_{21} & \dots & \varepsilon_{2T} \\ \dots & \dots & \dots & \dots \\ \pi_v & \varepsilon_{v1} & \dots & \varepsilon_{vT} \\ \dots & \dots & \dots & \dots \\ \pi_N & \varepsilon_{N1} & \dots & \varepsilon_{NT} \end{pmatrix}$

=

$H'$

+

$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

=

$H'$

Beide Formen haben gewisse Vorteile. Während die Wachstumskurvenanalyse und die Tests im allgemeinen multivariaten Modell meist in übersichtlicher Matrixform (5.7) formuliert werden (Timm, 1975; Jöreskog, 1979; Morrison, 1976), wird die ausgerollte Form bei der univariaten Varianzanalyse und bei „Pooling Cross Sections with Times Series Analysis“ verwendet. Zudem ist die ausgerollte Form flexibler, etwa bei Berücksichtigung von „missing data“ für verschiedene Zeitpunkte bei verschiedenen Personen. Wir werden in der folgenden Darstellung meistens die „ausgerollte Form“ verwenden. Als allgemeine Operation für das Ausrollen (also aus einer Matrix einen Vektor zu machen) wurde eingeführt (z.B. McDonald & Swaminathan, 1973):

$$(5.8) \quad \text{vec} \begin{pmatrix} Y_{11} & \dots & Y_{1T} \\ \dots & \dots & \dots \\ Y_{N1} & \dots & Y_{NT} \end{pmatrix} = \begin{bmatrix} Y_{11} \\ \dots \\ Y_{1T} \\ \dots \\ Y_{N1} \\ \dots \\ Y_{NT} \end{bmatrix} \quad \begin{array}{l} \text{mit der Regel:} \\ \text{vec}(ABC) = A \otimes C' \text{ vec}(B) \end{array}$$

Unter den Annahmen (5.1) bis (5.5) erhält man für die Varianzen von  $Y_{vt}$ :

$$\text{Var}(Y_{vt}) = \text{Var}(\pi_v + \epsilon_{vt}) = \sigma_\pi^2 + \sigma_\epsilon^2$$

und für die Kovarianzen von  $Y_{vt}$  und  $Y_{vt'}$ :

$$\text{Cov}(Y_{vt}, Y_{vt'}) = \sigma_\pi^2 \text{ für } t \neq t'$$

Die Matrixform als Varianz-Kovarianzmatrix für die v-Person ist:

$$(5.9a) \quad \text{cov} \begin{pmatrix} Y_{v1} \\ Y_{v2} \\ \vdots \\ Y_{vT} \end{pmatrix} = \underbrace{\begin{pmatrix} \sigma_\pi^2 & \sigma_\pi^2 & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2 & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2 & \sigma_\pi^2 \end{pmatrix}}_T + \underbrace{\begin{pmatrix} \sigma_\epsilon^2 & 0 & 0 \\ 0 & \sigma_\epsilon^2 & 0 \\ 0 & 0 & \sigma_\epsilon^2 \end{pmatrix}}_T = \underbrace{\begin{pmatrix} \sigma_\pi^2 + \sigma_\epsilon^2 & \sigma_\pi^2 & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2 + \sigma_\epsilon^2 & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2 & \sigma_\pi^2 + \sigma_\epsilon^2 \end{pmatrix}}_T \cdot T$$

$$(5.9b) \quad = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \cdot (\sigma_\pi^2 + \sigma_\epsilon^2) \quad \begin{array}{l} \text{wobei} \\ \rho =: \frac{\sigma_\pi^2}{\sigma_\pi^2 + \sigma_\epsilon^2} \end{array} \quad \left[ = \begin{array}{l} \text{Intraclass*} \\ \text{Korrelation} \end{array} \right]$$

In dieser Form der Varianz-Kovarianz-Matrix werden die speziellen Eigenschaften sichtbar:

1. Homoskedastizität über die Zeitpunkte hinweg
2. Die spezielle Symmetrie (compound symmetry), die einerseits durch die Unkorreliertheit der Störglieder zwischen verschiedenen Zeitpunkten, andererseits durch die konstante Korrelation benachbarter oder weit auseinander liegender Messungen bedingt ist. Dabei entsteht die Korrelation vermittels des Personenfaktors.

Diese beiden Annahmen sind simultan testbar über den von Wilks (1932) vorgeschlagenen Likelihoodratiotest, für den Box (1949) approximative  $\chi^2$ - und F-Verteilungen vorgeschlagen hat (Morrison, 1976, S. 250). Huynh & Feldt (1970) haben nachgewiesen, daß für bestimmte Tests im Rahmen dieses Modells etwas weniger restriktive Annahmen bezüglich der Kovarianzen nötig sind. Greenhouse & Geisser (1959) haben approximative Grenzen gefunden für „konservative“ Tests, die bei nicht allzu großer Modellverletzung für das Testen der Mittelwerthypothesen verwendet werden können (siehe auch Morrison, S. 214ff.). Diese Überlegungen zeigen, daß Tests im Rahmen des „Mixed“-Modells auch bei Modellverletzungen gerade noch tragbar sind.

### *Kritik an den Annahmen:*

Die durch  $\varepsilon_{vt}$  repräsentierten Störglieder stellen selbst Einflüsse dar, die einer Vielfalt von Prozessen unterworfen sein können. Gerade bei Beobachtungen über die Zeit ist es naheliegend, Prozesse zu vermuten, die aufgrund ihrer zeitlichen Kontinuität Kovarianzen zwischen den Störgliedern erzeugen. Formalisiert werden solche Prozesse meist durch autoregressive bzw. Moving-Average-Prozesse (siehe Box & Jenkins (1970)). Beide Arten der Prozesse in den Störgliedern produzieren eine Kovarianzstruktur in den  $\varepsilon_{vt}$  ( $t = 1, \dots, T$ ), in der nebeneinander liegende Zeitpunkte höher korrelieren als weiter auseinander liegende. Das würde bei  $T=4$  zu folgender Form von Kovarianzstruktur zwischen  $\varepsilon_{vt}$  führen:

$$(5.10) \quad \text{cov} \begin{pmatrix} \varepsilon_{v1} \\ \varepsilon_{v2} \\ \varepsilon_{v3} \\ \varepsilon_{v4} \end{pmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix} \cdot \sigma_\varepsilon^2 \quad \text{wobei} \quad 1 \geq \rho_1 \geq \rho_2 \geq \rho_3$$

Auch die Annahme der Homoskedastizität ist bei manchen Anwendungen zu restriktiv, da die Variabilität eines Prozesses im Laufe der Zeit zu- bzw. abnehmen kann. Das könnte selbst ein interessanter Aspekt eines Prozesses sein und sollte daher nicht von vornherein fälschlicherweise durch zu restriktive Annahmen ausgeschlossen werden.

Eine generelle Alternative\*\* zu den Restriktionen bezüglich E ist, die Kovarianzmatrix *völlig* ohne Restriktionen zu belassen

\* Siehe Winer (1971) und Nerlove (1971a,b)

\*\* Neben dieser Alternative gibt es auch die Möglichkeit, spezielle Annahmen für  $\varepsilon$  (z.B. ARIMA) und damit für die Kovarianzen von  $\varepsilon$  zu wählen (siehe z.B. Jöreskog (1979)).

$$(5.11) \quad \text{cov} \begin{pmatrix} \epsilon_{v1} \\ \epsilon_{v2} \\ \vdots \\ \epsilon_{vT} \end{pmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{21} & \sigma_{1T} \\ \sigma_{21} & \sigma_{22} & \sigma_{2T} \\ \vdots & \vdots & \vdots \\ \sigma_{T1} & \sigma_{T2} & \sigma_{TT} \end{bmatrix} = \Sigma_{\epsilon}$$

Für die Kovarianzmatrix von  $Y_{vt}$  ergibt sich dann noch:

$$(5.12) \quad \text{cov} \begin{pmatrix} Y_{v1} \\ Y_{v2} \\ \vdots \\ Y_{vT} \end{pmatrix} = \text{cov} \begin{pmatrix} \pi_v + \epsilon_{v1} \\ \pi_v + \epsilon_{v2} \\ \vdots \\ \pi_v + \epsilon_{vT} \end{pmatrix} = \begin{bmatrix} \sigma_{\pi}^2 & \sigma_{\pi}^2 & \sigma_{\pi}^2 \\ \sigma_{\pi}^2 & \sigma_{\pi}^2 & \sigma_{\pi}^2 \\ \vdots & \vdots & \vdots \\ \sigma_{\pi}^2 & \sigma_{\pi}^2 & \sigma_{\pi}^2 \end{bmatrix} + \Sigma_{\epsilon} =: \Sigma_e$$

wobei mit  $e$  die Summe der Elemente:  $e_{vt} = \pi_v + \epsilon_{vt}$  abgekürzt werden soll.

Diese erweiterte Annahme trifft man in den sogenannten „multivariaten“ Modellen an, in denen die  $T$  Messungen der *einen* inhaltlichen Variablen als  $T$  Variable aufgefaßt werden (siehe Kap. 3.2 und Bock 1963, 1975, 1979). Voraussetzung ist allerdings, daß das  $N$  so groß ist, daß die Tests der multivariaten Hypothesen noch durchgeführt werden können.

### 5.1.2 Zur Identifikation und Interpretation der Effektparameter

Man nimmt an, daß die Verteilung der abhängigen Variablen durch die Erwartungswerte:  $E(y_v)$  und durch die Kovarianzen  $\text{Cov}(y_v)$  beschrieben werden können. Andererseits wird unterstellt, daß die  $Y$ -Werte durch das Modell bestimmt sind. Die Frage der Identifikation bezieht sich darauf, inwiefern aus den Erwartungswerten und den Kovarianzen von  $y$  die Modellparameter  $\mu$ ,  $\tau_1, \dots, \tau_T$  und  $\Sigma_e$  eindeutig berechnet werden können. Dabei bezieht sich die Identifikation auf die Population und deren Parameter, nicht aber auf die Stichprobenschätzungen.

Da gilt, daß die Kovarianzmatrix der  $Y$ -Werte gleich  $\Sigma_e$  ist (siehe 5.12), können diese Modellparameter eindeutig berechnet werden. Man kann jedoch im allgemeinen multivariaten Ansatz die Parameter  $\sigma_{\pi}^2$  und die Matrix der Kovarianzen  $\Sigma_e$  nicht eindeutig bestimmen. Dies wäre für den restringierten Fall (Compound Symmetry) möglich. Dort gibt es die Gleichungen:

$$(5.13) \quad \text{Var}(Y_{vt}) = \sigma_{\pi}^2 + \sigma_{\epsilon}^2 \text{ und } \text{Cov}(Y_{vt}, Y_{vt'}) = \sigma_{\pi}^2$$

so daß  $\sigma_{\pi}^2$  und  $\sigma_{\epsilon}^2$  eindeutig berechenbar sind.

Aus den Erwartungswerten kann man versuchen, die übrigen Modellparameter:  $\mu$ ,  $\tau_1$ ,  $\tau_2$ ,  $\dots$ ,  $\tau_T$  aus den Gleichungen (5.14) herzuleiten:

$$(5.14) \quad E \begin{pmatrix} Y_{v1} \\ \dots \\ Y_{vT} \end{pmatrix} =: \begin{pmatrix} \mu_1 \\ \dots \\ \mu_T \end{pmatrix} = \begin{pmatrix} \mathbf{1} & 1 \dots 0 \\ & \dots \\ & 0 \dots 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \dots \\ \tau_T \end{pmatrix} = \begin{pmatrix} \mu + \tau_1 \\ \dots \\ \mu + \tau_T \end{pmatrix}$$

$$\mu = H \quad \tau$$

Die restlichen Summanden fallen weg, da zu den Modellannahmen (5.2, 5.5) gehört, daß die Erwartungswerte der  $\pi$ 's und  $\varepsilon$ 's Null sind. Man sieht, daß man  $T + 1$  Modellparameter aus den  $T$  Erwartungswerten nicht eindeutig berechnen kann.

### *Restriktionen und schätzbare Funktionen:*

Um die Eindeutigkeit herzustellen, könnte man einfach zusätzlich die Restriktionen einführen, daß  $\mu$  oder  $\sum_{t=1}^T \tau_t$  oder  $\tau_1$  gleich Null ist. Dadurch könnte man jeweils alle Modellparameter eindeutig aus den Erwartungswerten berechnen. Allerdings müßten für solche Annahmen Begründungen geliefert werden.

Ein anderer Weg läge nicht im Interesse an Einzelparametern, sondern an speziellen, eindeutig bestimmbar Linearkombinationen der Parameter („schätzbare Funktionen“, s.a. Scheffè, 1959).

Im folgenden sollen einige speziell im Rahmen von Zeitsequenzen interessierende Typen solcher Funktionen betrachtet werden. Diese Funktionen in den „alten Parametern“  $\mu$ ,  $\tau_t$  werden dann jeweils als neue Parameter bezeichnet. Im allgemeinen sollen diese der Einfachheit halber als Funktionen in den Erwartungswerten  $\mu_t$  geschrieben werden. Das hat darüber hinaus den Vorteil, daß es sich dann auf jeden Fall um schätzbare Funktionen handeln muß. Diese Typen werden anhand von Beispielen eingeführt.

#### *Typ 1: Abweichungskontraste*

Beispiel: Es werde nach der 1. Messung des Gewichts eine Behandlung durchgeführt. Insgesamt werden 3 Messungen vorgenommen ( $T = 3$ ).

Dabei könnte man sich für die Unterschiede zwischen der 1. und der  $t$ . Messung interessieren. Zusätzlich soll ein Parameter für das generelle Niveau eingeführt werden. Dann erhält man für die neuen Parameter folgende Funktionen in den „alten“ Parametern bzw. in den Erwartungswerten:

$$\gamma_0 = \frac{\mu_1 + \mu_2 + \mu_3}{3} = \mu + \frac{\tau_1 + \tau_2 + \tau_3}{3}$$

$$\gamma_1 = \mu_1 - \mu_2 = \tau_1 - \tau_2$$

$$\gamma_2 = \mu_1 - \mu_3 = \tau_1 - \tau_3$$

oder in Matrixform:

$$(5.15) \quad \begin{matrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{matrix} = \begin{matrix} 1/3 & 1/3 & 1/3 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{matrix} \cdot \begin{matrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{matrix} = \begin{matrix} 1 & 1/3 & 1/3 & 1/3 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{matrix} \begin{matrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{matrix}$$

neue
Erwar-
alte  
Parameter
tungswerte
Parameter

Diese Linearkombinationen in den Modellparametern sind im Rahmen dieses Modells eindeutig, da sie auch zugleich als Linearkombinationen in den 1. Momenten dargestellt sind.

### Typ 2: Helmertkontraste

Bei gleicher Fragestellung könnte eine andere Differenzbildung folgende sein:

Neben dem generellen Niveau:  $\gamma_0 = \frac{1}{T}(\mu_1 + \mu_2 \dots + \mu_T)$  interessiert der

Unterschied zwischen der ersten und dem Mittel der restlichen Messungen, der Unterschied zwischen der zweiten und dem Mittel der restlichen Messungen usw. (siehe Bock, 1975, Finn, 1974). Für das obige Beispiel ist bei  $T=3$

$$\begin{aligned} \gamma_0 &= \frac{1}{3} (\mu_1 + \mu_2 + \mu_3) = \mu + \frac{1}{3} (\tau_1 + \tau_2 + \tau_3) \\ \gamma_1 &= \mu_1 - \frac{1}{2} (\mu_2 + \mu_3) = \tau_1 - \frac{1}{2} (\tau_2 + \tau_3) \\ \gamma_2 &= \mu_2 - \mu_3 = \tau_2 - \tau_3 \end{aligned}$$

in Matrixschreibweise:

$$(5.16) \quad \begin{matrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{matrix} = \begin{matrix} 1/3 & 1/3 & 1/3 \\ 1 & -1/2 & -1/2 \\ 0 & 1 & -1 \end{matrix} \begin{matrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{matrix} = \begin{matrix} 1 & 1/3 & 1/3 & 1/3 \\ 0 & 1 & -1/2 & -1/2 \\ 0 & 0 & 1 & -1 \end{matrix} \begin{matrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{matrix}$$

### Typ 3: Polynomkontraste

Beispiel: Im Rahmen einer Entwicklungsstudie könnte der zeitliche Verlauf der intellektuellen Entwicklung Untersuchungsgegenstand sein: Ist neben der generellen Höhe ein linearer, quadratischer, kubischer usw. Verlauf zu berücksichtigen?

$$(5.17) \quad \mu_t = \gamma_0 t^0 + \gamma_1 t + \dots + \gamma_p t^p; \quad \begin{array}{l} \gamma_0 \text{ gibt den Beitrag der Konstanten an,} \\ \gamma_1 \text{ gibt den Beitrag der Geraden an,} \\ \text{u.s.w., die in } \mu_t \text{ nachweisbar sind} \end{array}$$

Für 3 Zeitpunkte (Zeit = 1,2,3):

$$\begin{aligned} \mu_1 &= \mu + \tau_1 = \gamma_0 + \gamma_1 + \gamma_2 \\ \mu_2 &= \mu + \tau_2 = \gamma_0 + \gamma_1 2 + \gamma_2 4 \\ \mu_3 &= \mu + \tau_3 = \gamma_0 + \gamma_1 3 + \gamma_2 9 \end{aligned}$$

in Matrixform:

$$(5.18) \quad \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} = \begin{matrix} t^0 & t^1 & t^2 \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \end{matrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} \quad \begin{array}{l} K \text{ ist eine} \\ \text{Vandermodematrix} \\ \text{(s. Timm, 1975)} \end{array}$$

Diese Spezifikation ist *vollständig*, da bei 3 Zeitpunkten ein quadratischer Trend nachweisbar ist. Für T Zeitpunkte kann man immer ein Polynom T-1. Ordnung anpassen. Aus inhaltlichen Gründen könnte dagegen bekannt sein, daß zwar für T Zeitpunkte Messungen vorliegen, aber ein Polynom T-j. Ordnung ( $j > 1$ ) zur Beschreibung des Trends ausreicht (bei 3 Zeitpunkten soll neben der Konstanten der lineare Trend genügen). Eine solche Spezifikation heißt *unvollständig*.

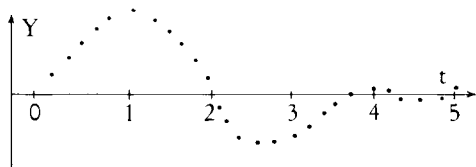
Z.B.

$$(5.19) \quad \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix}$$

### Typ 4: Fourier-Kontraste

Bei Trends, die als periodische Schwankungen darstellbar sind, sind Polynome keine optimale Form der Darstellung. Dann ist es naheliegend, auf Sinus- (bzw. Cosinus) Funktionen zurückzugreifen.

Beispiel: Nach Darbietung eines Reizes werden EEG-Ströme gemessen  $Y$ . Dabei könnte man etwa folgenden Verlauf feststellen, der durch Sinusfunktionen gut beschreibbar wäre:



Es wird versucht, den Trend in Sinusschwingungen mit unterschiedlicher Frequenz und Phasenverschiebung zu zerlegen:

$$(5.20) \quad \mu_t = \gamma_0 + C_1 \cdot \sin(\alpha_1 t + \varphi_1) + C_2 \cdot \sin(\alpha_2 t + \varphi_2)$$

Hier sind zwar nur 2 unterschiedliche Sinusschwingungen enthalten, jedoch ist dieser Ansatz generell für mehrere Frequenzen (z. B.: Anderson, 1971; Fuller, 1976) erweiterbar.

$\alpha_1, \alpha_2$ : 2 verschiedene Frequenzen

$\varphi_1, \varphi_2$ : die dazugehörigen Phasenverschiebungen der Sinusfunktion (Bei  $t=0$  hat die Sinusfunktion den Wert des Phasenwinkels  $\varphi_i$ ).

Die Größen  $\alpha$  und  $\varphi$  werden in Grad oder Bogenmaß angegeben.

$C_1, C_2$ : die Koeffizienten der Amplituden messen die Stärke des Variierens der Sinusfunktionen der entsprechenden Frequenz.

Da versucht werden soll,  $\mu_t$  in lineare Funktionen zu zerlegen, kann man mit der obigen Darstellung noch nicht zufrieden sein, da Koeffizienten  $\alpha_i$  und  $\varphi_i$  vorkommen, von denen der Sinus zu berechnen ist; ein weiterer Schritt zu einer Linearisierung kann durch folgende Umformung gemacht werden:

$$(5.21) \quad \begin{aligned} C_1 \sin(\alpha_1 t + \varphi_1) &= C_1 \sin(\alpha_1 t) \cdot \cos \varphi_1 + C_1 \cdot \cos(\alpha_1 t) \cdot \sin \varphi_1 \\ &= \gamma_2 \sin(\alpha_1 t) + \gamma_1 \cos(\alpha_1 t) \end{aligned}$$

$$(5.22) \quad \text{mit } \gamma_2 = C_1 \cos \varphi_1 \text{ und } \gamma_1 = C_1 \sin \varphi_1$$

Entsprechend kann man auch den 3. Summanden von (5.20) zerlegen, so daß man

$$(5.23) \quad \mu_t = \gamma_0 + \gamma_1 \cos(\alpha_1 t) + \gamma_2 \sin(\alpha_1 t) + \gamma_3 \cos(\alpha_2 t) + \gamma_4 \sin(\alpha_2 t)$$

erhält.

Wären  $\alpha_1$  und  $\alpha_2$  bekannt, könnte man dieses Gleichungssystem zur Berechnung der  $\gamma_i$  mit  $\sin(\alpha_i t)$ ,  $\cos(\alpha_i t)$  als Bekannten verwenden. Aus den Koeffizienten  $\gamma_i$  könnte man dann mit Hilfe der Gleichungen (5.22) die Amplituden  $C_1$  und die Phasen  $\varphi_1$  errechnen.



Eine Möglichkeit der Bestimmung der  $\alpha_1$  läge in deren systematischen Variation (siehe Bloomfield, 1976). Andererseits zeigt die Fourieranalyse (siehe Fuller und Bloomfield), daß bei endlich vielen Meßzeitpunkten sogar jede Funktion durch eine Summe von Sinus- und Cosinusfunktionen (ähnlich 5.23) dargestellt werden kann, falls genauso viele Koeffizienten  $\gamma_i$  berücksichtigt werden wie Zeitpunkte vorliegen. Dabei kann man für die Frequenzen  $\alpha_i$  folgende einfache Form verwenden:

$$(5.24a) \quad \alpha_i = \frac{2 \cdot \pi}{T} i \text{ (im Bogenmaß); } i = 1, \dots, L$$

$$(5.24b) \quad = \frac{360}{T} \cdot i \text{ (in Grad) wobei } L \text{ größte ganze Zahl ist, die kleiner (gleich) } T/2 \text{ ist.}$$

Sind in obigem Beispiel 5 Messungen bekannt, ist  $L = 2$  (da 2 die größte ganze Zahl, die kleiner als  $5/2$  ist). Die beiden  $a_i$  sind gemäß (5.24b):

$$\alpha_1 = \frac{360^\circ}{5} \cdot 1 = 72^\circ; \alpha_2 = \frac{360^\circ}{5} \cdot 2 = 144^\circ$$

In (5.23) eingesetzt erhält man daher für  $\mu_t$ :

$$(5.25) \quad \mu_t = \gamma_0 + \gamma_1 \cos(72 \cdot t) + \gamma_2 \sin(72 \cdot t) + \gamma_3 \cos(144 \cdot t) + \gamma_4 \sin(144 \cdot t)$$

Die Sinus- und Cosinuswerte sind unmittelbar als feste Größen berechenbar. Die Matrixgleichung ist für  $T = 5$

$$(5.26) \quad \begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & .31 & .95 & -.81 & .59 \\ 1 & -.81 & .59 & .31 & -.95 \\ 1 & -.81 & -.59 & .31 & .95 \\ 1 & .31 & -.95 & -.81 & -.59 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{bmatrix}$$

Kon    Cos    Sin    Cos    Sin  
stante (72 · t) (72 · t) (144 · t) (144 · t)

(höhere Frequenz = schnellere Schwingungen)

Man beachte, daß die Summe der Produkte der Elemente aus je zwei verschiedenen Spalten Null ergibt.

Denn es gilt allgemein:

$$(5.27a) \quad \sum_{t=0}^{T-1} \cos(\alpha_i t) \cdot \sin(\alpha_j t) = 0 \quad \text{gemischt: Sinus und Kosinusprodukte}$$

$$(5.27b) \quad \sum_{t=0}^{T-1} \cos(\alpha_i t) \cos(\alpha_j t) = 0 \quad (j \neq i) \quad \begin{array}{l} \text{Kosinusprodukte mit} \\ \text{unterschiedlicher} \\ \text{Frequenz} \end{array}$$

$$(5.27c) \quad \sum_{t=0}^{T-1} \sin(\alpha_i t) \sin(\alpha_j t) = 0 \quad (j \neq i) \quad \begin{array}{l} \text{Sinusprodukte mit} \\ \text{unterschiedlicher} \\ \text{Frequenz} \end{array}$$

$$(5.27d) \quad \sum_{t=0}^{T-1} \sin^2(\alpha_i t) = \sum_{t=0}^{T-1} \cos^2(\alpha_i t) = \frac{T}{2} \quad \text{für } i \neq 0 \quad i \neq T/2$$

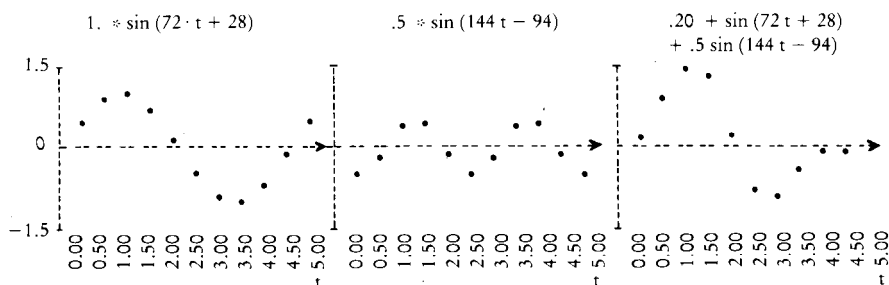
Für ein konkretes Zahlenbeispiel sei  $C_1 = 1$  und  $C_2 = .5$ , für  $\gamma_0 = .20$ ,  $\varphi_1 = 28^\circ$ ,  $\varphi_2 = -94^\circ$ .

Die  $\gamma_i$  ( $i=1 \dots 4$ ) ergeben sich aus Formel (5.22):  $\gamma_1 = .47$ ,  $\gamma_2 = .883$   
 $\gamma_3 = -.50$ ,  $\gamma_4 = -.035$

Man erhält

$$\text{für die } \mu's: \quad \begin{array}{c|c|c|c|c} \mu_0 & \mu_1 & \mu_2 & \mu_3 & \mu_4 \\ \hline .17 & 1.56 & .22 & -.88 & -.07 \end{array}$$

Der Übersicht halber wurden die einzelnen Komponenten und die Gesamtfunktion geplottet (mit Zwischenpunkten):



### Einsetzung der Reparametrisierungstypen

Durch die 4 Typen von Kontrasten wurden jeweils neue Parameter  $\gamma$  eingeführt. Das Modell (5.6) werde nun in den neuen Parameter angesetzt. Dabei ist es nur notwendig, für die  $v$ -te Person in die alte Parametrisierung:

$$(5.28a) \quad \gamma_v = H\tau + e_v \quad \text{bzw.}$$

$$(5.28b) \quad \gamma_v = \mu + e_v \quad \text{die neuen Parameter } \gamma \text{ einzusetzen.}$$

Für die Typen 1 und 2 ist  $\gamma$  als Linearkombination von  $\mu$  bzw.  $H\tau$  geschrieben:

$$(5.29) \quad \gamma = R\mu = RH\tau$$

Da aber  $\mu$  ersetzt werden soll, muß  $\mu$  vorweg noch als Funktion von  $\gamma$  dargestellt werden:

Es sei  $R$  eine Matrix, deren Inverse existiert. Dann erhält man:

$$(5.30) \quad R^{-1}\gamma = \mu = H\tau$$

Somit lautet das reparametrisierte Modell für die  $v$ te Person

$$(5.31) \quad y_v = R^{-1}\gamma + e_v$$

In Beispiel 1 ist:

$$R\mu = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \quad \mu = R^{-1} \cdot \gamma = \begin{bmatrix} 1 & 1/3 & 1/3 \\ 1 & -2/3 & 1/3 \\ 1 & 1/3 & -2/3 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix}$$

In Beispiel 2 ist:

$$R\mu = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1 & -1/2 & -1/2 \\ 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \quad \mu = R^{-1} \cdot \gamma = \begin{bmatrix} 1 & 2/3 & 0 \\ 1 & -1/3 & 1/2 \\ 1 & -1/3 & -1/2 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix}$$

Die Matrix  $R^{-1}$  stellt eine Basismatrix der neuen Parametrisierung dar. Sie wird meist mit  $K$  abgekürzt (siehe Bock, 1975, 1979). Daher kann man Gleichung (5.31) auch so schreiben:

$$(5.32) \quad y_v = K\gamma + e_v$$

Bei den Typen 3 und 4 wurde  $\mu$  von vornherein als Funktion von neuen Variablen aufgefaßt, deren Stärke durch den Parametervektor  $\gamma$  untersucht werden soll, so daß die Inverse von  $R$  nicht zu berechnen ist; die Einsetzung kann ohne Umrechnung unmittelbar vorgenommen werden.

Bei Beispiel 3 ist:  $K = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}$  bei vollständiger Spezifikation des Trends

Bei Beispiel 4 ist  $K$  durch die Matrix der  $\cos(\alpha_i \cdot t)$  bzw.  $\sin(\alpha_i \cdot t)$  gegeben.

Auch bei diesem Beispiel wäre es leicht möglich, eine unvollständige Spezifikation vorzunehmen, falls man aus inhaltlichen Gründen höhere Frequenzen weglassen könnte.

### *Orthogonale Reparametrisierung:*

Aus der Regressionsanalyse ist bekannt, daß die Interpretation der Parameter einfach ist, wenn die unabhängigen Variablen nicht miteinander korrelieren. Zu- oder Wegnahme eines Prädiktors verursacht keine Veränderung der Regressionskoeffizienten.

Diesen Vorteil versucht man auch bei der orthogonalen Reparametrisierung zu nutzen. Sie liegt dann vor, wenn gilt:

$$K'K = \text{Diagonalmatrix}$$

Diese Eigenschaft trifft für die Helmertkontraste, wegen (5.27) für die Fourierkontraste und für die orthogonalen Polynome (siehe auch Anderson (1971), Bock (1975)) zu.

Beispiel: bei  $T = 3$  lautet die  $K$ -Matrix bei orthogonalen Polynomen:

$$K = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{weitere Beispiele kann man z.B. in Bock (1975) tabelliert finden}$$

## 5.2 Berücksichtigung von gruppenspezifischen Faktoren

$$(T \geq 2, G \geq 2, N \geq G, N \geq T)$$

Soll im Modell berücksichtigt werden, daß für unterschiedliche Gruppen von Personen spezifische Gruppenhaupt- bzw. Veränderungseffekte zu erwarten sind (Interaktion zwischen Zeit und Gruppen, wobei jede der Gruppen  $N_1, \dots, N_G$  Personen enthält), ist folgendes Modell üblich:::

$$Y_{vit} = \mu + \alpha_i + \pi_{v(i)} + \tau_t + (\alpha\tau)_{it} + \varepsilon_{vit}$$

Gegenüber dem Modell (5.1) sind hier Parameter für die Gruppenhaupteffekte ( $\alpha_i, i = 1, \dots, G$ ) und für die Interaktion zwischen Gruppen und Zeitpunkte

---

\* Siehe Winer (1971), Bock (1975), Finn (1969). Finn trennt in eine Interaktion zwischen Zeit  $x$  Personeneffekt und ein Störglied. Da dieser Interaktionseffekt nur schätzbar ist, wenn jede Person zu jedem Zeitpunkt mehrmals gemessen wird, lassen wir ihn weg.

$(\alpha\tau)_{it}$  enthalten. Für  $\pi_{v(i)}$  gelte:  $v$  variiert zwischen 1 bis  $N$ . Das  $(i)$  soll angeben, zu welcher Gruppe die  $v$ -te Person gehört. Zudem seien die Personeneffekte wiederum Zufallsgrößen mit den Annahmen (5.3 - 5.5). Für  $\varepsilon_{vt}$  gelten die gleichen Überlegungen wie in Kapitel 5.1.1.

Mit diesen Annahmen erhält man als Erwartungswerte der  $Y_{vit}$ :

$$(5.33) \quad E(Y_{vit}) = \mu_{it} = \mu + \alpha_i + \tau_t + (\alpha\tau)_{it}$$

Die Anzahl der Modellparameter  $(\mu, \alpha_i, \tau_t, (\alpha\tau)_{it})$  ist:  $1 + G + T + G \cdot T$ . Diese lassen sich nicht eindeutig aus den  $G \cdot T$  Erwartungswerten  $\mu_{it}$  herleiten. Es liegt ein Identifikationsproblem vor.

Es soll daher versucht werden, eine Reparametrisierung zu finden, die höchstens  $G \times T$  schätzbare lineare Funktionen als neue Parameter enthält. Wir wollen einen Weg wählen, der von den  $\mu_{it}$  ausgeht, die als bekannt vorausgesetzt werden dürfen. Ferner arbeiten wir mit Kontrasttypen, die schon für den Fall ohne Gruppenfaktor vorgestellt wurden (gruppenspezifische Parametrisierung). In einem weiteren Schritt sollen dann diese neuen Parameter für Gruppenvergleiche noch weiter zerlegt werden (gruppenübergreifende Parametrisierung). Diese Überlegungen lassen sich wiederum an einem Beispiel gut veranschaulichen.

*Beispiel:* Es seien 2 Gruppen ( $G = 2$ ), die zu 3 Zeitpunkten ( $T = 3$ ) gemessen wurden, gegeben. Das Modell kann geschrieben werden als:

$$Y_{vit} = \mu_{it} + \pi_{v(i)} + \varepsilon_{vt}$$

Die Erwartungswerte für die 1. Gruppe sollen dargestellt werden als orthogonale Polynome:

$$(5.34) \quad \begin{array}{|c|} \hline \mu_1 \\ \hline \end{array} = \begin{array}{|c|} \hline \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \hline \end{array} = \begin{array}{|ccc|} \hline 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \\ \hline \end{array} \cdot \begin{array}{|c|} \hline b_{11} \\ b_{12} \\ b_{13} \\ \hline \end{array}$$

$= K \cdot b_1$

wobei  $b_{11}$  das generelle Niveau  
 $b_{12}$  den zusätzlichen linearen Trend  
 $b_{13}$  den zusätzlichen quadratischen Trend  
für die 1. Gruppe als neue Parameter repräsentieren

Für die 2. Gruppe ist:

$$(5.35) \quad \begin{array}{|c|} \hline \mu_2 \\ \hline \end{array} = \begin{array}{|c|} \hline \mu_{21} \\ \mu_{22} \\ \mu_{23} \\ \hline \end{array} = \begin{array}{|ccc|} \hline 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \\ \hline \end{array} \cdot \begin{array}{|c|} \hline b_{21} \\ b_{22} \\ b_{23} \\ \hline \end{array}$$

$= K \cdot b_2$

Die Parameter  $b_{21}, b_{22}, b_{23}$  stellen für die 2. Gruppe die neuen Parameter dar.

Die Matrixform für beide Gruppen ist:

$$(5.36) \quad \begin{array}{c} \mu_1 \\ \hline \mu_2 \end{array} = \begin{array}{c} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \hline \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{array} = \begin{array}{|c|c|} \hline K & O \\ \hline \hline O & K \\ \hline \end{array} \begin{array}{c} b_{11} \\ b_{12} \\ b_{13} \\ \hline b_{21} \\ b_{22} \\ b_{23} \end{array} = ({}_2I_2 \otimes K) \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

$$\downarrow$$

$$\begin{array}{|c|c|} \hline 1 & 0 \\ \hline 0 & 1 \\ \hline \end{array} = {}_2I_2$$

wobei das Kroneckerprodukt:  $\otimes$  bedeutet:

$$(5.37) \quad A \otimes B = \begin{pmatrix} a_{11} \cdot B & \dots & a_{1m} \cdot B \\ \hline a_{n1} \cdot B & \dots & a_{nm} \cdot B \end{pmatrix}$$

Diese Art der gruppenspezifischen Reparametrisierung verwenden Jöreskog (1979) und Morrison (1976). Sie ist wegen ihrer Übersichtlichkeit leicht zu interpretieren.

In einem weiteren Schritt könnte man versuchen, die gruppenspezifischen Trendparameter in solche umzuformen, die einerseits einen generellen gruppenunspezifischen Trend, andererseits gruppenspezifische Trendunterschiede repräsentieren. Manche dieser so gewonnenen Parameter können dann wiederum als Haupteffekte - andere als Interaktionsparameter - interpretiert werden.

### Fortsetzung der Beispiels:

Gesucht sind Parameter, die die folgenden Eigenschaften haben:\*

Durchschnitt des	Niveaus	$\begin{array}{ c } \hline b_{11}^* \\ b_{12}^* \\ b_{13}^* \\ \hline b_{21}^* \\ b_{22}^* \\ b_{23}^* \\ \hline \end{array}$	$=$	$\begin{array}{ c } \hline b_1^* \\ \hline b_2^* \\ \hline \end{array}$	$= 1/2$	$\begin{array}{ c } \hline b_{11} + b_{21} \\ b_{12} + b_{22} \\ b_{13} + b_{23} \\ \hline b_{11} - b_{21} \\ b_{12} - b_{22} \\ b_{13} - b_{23} \\ \hline \end{array}$	$=$
	linearen Trends						
	quadratischen Trends						
Gruppenunterschiede im	Niveau						
	linearen Trend						
	quadratischen Trend						

\* Es ergibt sich folgende Interpretation von  $b^*$  in der Terminologie von Haupt- und Interaktionsparametern (siehe Bock (1975))

Durchschnitt im Niveau	}	Haupteffekte (Zeit)
Durchschnitt im linearen Trend		
Durchschnitt im quadrat. Trend		
Unterschied im Niveau	}	Haupteffekte (Gruppe) Interaktion (Zeit x Gruppe)
Unterschied im lin. Trend		
Unterschied im quadrat. Trend		

$$= \begin{bmatrix} 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & -1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & -1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & -1/2 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \\ b_{21} \\ b_{22} \\ b_{23} \end{bmatrix} = \begin{bmatrix} 1/2 I & 1/2 I \\ 1/2 I & -1/2 I \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix}$$

Die obige Gleichung bleibt sicher gültig bei Einführung der identischen Matrix:

$$I = \begin{bmatrix} I & O \\ O & I \end{bmatrix}$$

$$\begin{aligned} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} &= \begin{bmatrix} K & O \\ O & K \end{bmatrix} \cdot \begin{bmatrix} I & I \\ I & -I \end{bmatrix} \cdot \begin{bmatrix} 1/2 I & 1/2 I \\ 1/2 I & -1/2 I \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\ &\quad \text{Durch Ausrechnen} \quad \text{Durch Ausrechnen} \\ \mu &= \begin{bmatrix} K & K \\ K & -K \end{bmatrix} \cdot \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix} = \left[ \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes K \right]_{K_D} \cdot \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix} \\ &\quad \text{wegen Definition von } b^* \end{aligned}$$

Die Gruppenmittel lassen sich durch eine solche Umformung somit in den neuen Parametern  $b_1^*$  und  $b_2^*$  darstellen:

$$(5.38) \quad \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} K & K \\ K & -K \end{bmatrix} \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix}; \quad \begin{bmatrix} \mu_2 \end{bmatrix} = \begin{bmatrix} K & -K \end{bmatrix} \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix}$$

Für jedes  $\mu_{it}$  ist damit eine Linearkombination in den neuen Parametern gegeben; daher kann man in der Modellgleichung für  $\mu_1$  und  $\mu_2$  nach (5.38)  $b_1^*$  und  $b_2^*$  einsetzen.

Es seien im Beispiel 3 Männer und 2 Frauen zu 3 Zeitpunkten untersucht worden. Die Modellgleichungen sind in Matrixform (s. S. 361).

Die Matrix  $G$  hat pro Person eine Zeile. Die Spalte gibt an, aus welcher Gruppe die Person stammt. Sie vervielfacht die entsprechenden Zeilen von  $K_D$  in dem Ausmaß der Gruppengrößen, wobei  $K_D$  die Kontrastbasismatrix für die Gruppenvergleiche ist.

Im Rahmen dieses Beispiels wurde eine Form des Modells eingeführt, die ganz allgemein für verschiedenartige Designs verwendet werden kann. Dieser neue

♂ 1. Gruppe	1. P.	$Y_{111}$	$\mu_1$	+	$\pi_{1(1)} + \varepsilon_{11}$	=	Durch Ein- setzen von (38)	$K$	$K$	$b_1^*$	$\pi_{1(1)} + \varepsilon_{11}$	}	$e_1$		
		$Y_{112}$			$\pi_{1(1)} + \varepsilon_{12}$						$\pi_{1(1)} + \varepsilon_{12}$			$\pi_{1(1)} + \varepsilon_{13}$	
		$Y_{113}$			$\pi_{1(1)} + \varepsilon_{13}$						$\pi_{1(1)} + \varepsilon_{13}$				
	2. P.	$Y_{211}$	$\mu_1$	+	$\pi_{2(1)} + \varepsilon_{21}$	=	Durch Ein- setzen von (38)	$K$	$K$	$b_2^*$	$\pi_{2(1)} + \varepsilon_{21}$	}	$e_2$		
		$Y_{212}$			$\pi_{2(1)} + \varepsilon_{22}$						$\pi_{2(1)} + \varepsilon_{22}$			$\pi_{2(1)} + \varepsilon_{23}$	
		$Y_{213}$			$\pi_{2(1)} + \varepsilon_{23}$						$\pi_{2(1)} + \varepsilon_{23}$				
	3. P.	$Y_{311}$	$\mu_1$	+	$\pi_{3(1)} + \varepsilon_{31}$	=	Durch Ein- setzen von (38)	$K$	$K$	+	$\pi_{3(1)} + \varepsilon_{31}$	}	$e_3$		
		$Y_{312}$			$\pi_{3(1)} + \varepsilon_{32}$						$\pi_{3(1)} + \varepsilon_{32}$			$\pi_{3(1)} + \varepsilon_{33}$	
		$Y_{313}$			$\pi_{3(1)} + \varepsilon_{33}$						$\pi_{3(1)} + \varepsilon_{33}$				
	♀ 2. Gruppe	4. P.	$Y_{421}$	$\mu_2$	+	$\pi_{4(2)} + \varepsilon_{41}$	=	Durch Ein- setzen von (38)	$K$	$-K$		$\pi_{4(2)} + \varepsilon_{41}$	}	$e_4$	
			$Y_{422}$			$\pi_{4(2)} + \varepsilon_{42}$						$\pi_{4(2)} + \varepsilon_{42}$			$\pi_{4(2)} + \varepsilon_{43}$
			$Y_{423}$			$\pi_{4(2)} + \varepsilon_{43}$						$\pi_{4(2)} + \varepsilon_{43}$			
5. P.	$Y_{521}$	$\mu_2$	+	$\pi_{5(2)} + \varepsilon_{51}$	=	Durch Ein- setzen von (38)	$K$	$-K$		$\pi_{5(2)} + \varepsilon_{51}$	}	$e_5$			
	$Y_{522}$			$\pi_{5(2)} + \varepsilon_{52}$						$\pi_{5(2)} + \varepsilon_{52}$			$\pi_{5(2)} + \varepsilon_{53}$		
	$Y_{523}$			$\pi_{5(2)} + \varepsilon_{53}$						$\pi_{5(2)} + \varepsilon_{53}$					

$$= \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix}}_Y \otimes K \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Das Schreiben mit Hilfe des Kroneckerprodukts macht das wiederholte Schreiben der Kontrastmatrizen überflüssig.

$$= \left[ \underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}}_G \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_{K_1} \otimes K \right] \begin{bmatrix} b_1^* \\ b_2^* \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Durch das Einführen der Matrix  $G$  (Gruppenbeziehungsmatrix) kann man die Matrix  $K_1$ , die bei der Parametrisierung eingeführt wurde, wieder verwenden.

Ansatz zerlegt komplexe Designs in einfache (einfaktorielle) und setzt sie mit Hilfe von Kroneckerprodukten zum geforderten eventuell multifaktoriellen Design zusammen. Beispiele wurden von Zelen & Federer (1966) und von Bock (1975) für viele Designs gezeigt.

Aufgrund dieser Überlegungen kann man daher auch für den allgemeinen Fall das varianzanalytische Modell für wiederholte Messungen (wie im Beispiel) in ausgerollter Form darstellen (5.39).

Für  $X$  könnte man auch  $G \cdot K_D$  einsetzen, wobei  $K_D$  dann die Designmatrix für einen Faktor (wie im Beispiel) oder aber eine Designmatrix für ein mehrfaktorielles Design wäre (s. Bock, 1975).

G gäbe dann wieder durch Einsetzen in den entsprechenden Zeilen und Spalten für jede Person an, zu welcher Gruppe sie gehört. Andererseits könnte



(5.39)

$$\begin{array}{c}
 \begin{array}{c} 1 \\ \dots \\ T \end{array} \begin{array}{c} y_1 \\ \dots \\ y_2 \\ \dots \\ y_v \\ \dots \\ y_N \end{array} \\
 \begin{array}{c} 1 \\ \dots \\ T \end{array} \\
 \begin{array}{c} 1 \\ \dots \\ T \end{array}
 \end{array}
 =
 \begin{array}{c}
 \begin{array}{c} 1 \\ \dots \\ T \end{array} \begin{array}{c} y_1 \\ \dots \\ y_2 \\ \dots \\ y_v \\ \dots \\ y_N \end{array} \\
 \begin{array}{c} 1 \\ \dots \\ T \end{array} \\
 \begin{array}{c} 1 \\ \dots \\ T \end{array}
 \end{array}
 =
 {}_N X_I \otimes K$$

$$\begin{array}{c}
 \begin{array}{c} b_1 \\ \dots \\ b_2 \\ \dots \\ b_l \end{array} \\
 \begin{array}{c} p^* \\ p \\ p \end{array}
 \end{array}
 +
 \begin{array}{c}
 \begin{array}{c} e_1 \\ \dots \\ e_2 \\ \dots \\ e_v \\ \dots \\ e_N \end{array} \\
 \begin{array}{c} p^* \\ p \\ p \end{array}
 \end{array}$$

$$y = X \otimes K \quad b \quad + \quad e$$

man auch für  $K$  jede von den im vorigen Abschnitt besprochenen Typen einsetzen. Sie könnte sich darüber hinaus bei zusätzlichem Design in den wiederholten Messungen (z.B. Swaminathan & Algina, 1977) oder bei Mehrfachantworten (s.U.) selbst wieder aus einfachen Typen zusammensetzen.

Der Erwartungswert von  $y$  kann geschrieben werden als:

(5.40a)  $E(y) = (X \otimes K) b$

Die Kovarianz von  $y_v$  (für die  $v$ -te Person) wurde schon in (5.12) behandelt. Die Kovarianz für  $y$  (alle Personen) ist

(5.40b)

$$\text{cov} \left( \begin{array}{c} y_1 \\ \dots \\ y_N \end{array} \right) = \text{cov} \left( \begin{array}{c} e_1 \\ \dots \\ e_N \end{array} \right) = \begin{array}{ccc} \Sigma_e & \dots & O \\ \dots & \dots & \dots \\ O & \dots & \Sigma_e \end{array} = I_N \otimes \Sigma_e$$

\* Im vollständigen Modell ist  $p=T$  und im unvollständigen  $p < T$

Das obige „ausgerollte“ Gleichungssystem (5.39) läßt sich „engerollt“ in Matrixschreibweise formulieren:\*

$$\begin{array}{ccccccc}
 \begin{array}{c} 1 \\ \vdots \\ \vdots \\ N \end{array} & \boxed{\begin{array}{ccc} Y_{11} & \dots & Y_{1T} \\ & \dots & \\ Y_{N1} & \dots & Y_{NT} \end{array}} & \begin{array}{c} 1 \\ \vdots \\ \vdots \\ N \end{array} & = & \begin{array}{c} 1 \\ \vdots \\ \vdots \\ N \end{array} & \boxed{X} \\
 & \begin{array}{c} 1 \dots T \\ N Y_T \end{array} & & & & \begin{array}{c} 1 \dots I \\ N X_I \end{array} & \\
 & & & & & & B_p
 \end{array}
 \quad
 \begin{array}{ccccccc}
 \begin{array}{c} 1 \\ \vdots \\ \vdots \\ p \end{array} & \boxed{\begin{array}{ccc} b_{11} & \dots & b_{1p} \\ & \dots & \\ b_{11} & \dots & b_{1p} \end{array}} & \begin{array}{c} 1 \\ \vdots \\ \vdots \\ p \end{array} & \begin{array}{c} \boxed{K'} \\ 1 \dots T \end{array} & + & \begin{array}{c} 1 \\ \vdots \\ \vdots \\ N \end{array} & \boxed{\begin{array}{ccc} e_{11} & \dots & e_{1T} \\ & \dots & \\ e_{N1} & \dots & e_{NT} \end{array}} \\
 & \begin{array}{c} 1 \dots p \\ B_p \end{array} & & & Q_T & + & \begin{array}{c} 1 \dots T \\ N E_T \end{array}
 \end{array}$$

Der Erwartungswert ist:

$$E(Y) = {}_N X_I B_p Q_T$$

In dieser Weise wird meist das Wachstumskurvenmodell formuliert (siehe z.B. Timm (1975), Grizzle & Allen (1969), Morrison (1976)).\*\*

Wegen der Einheitlichkeit der Terminologie formulieren wir das Modell auch als

$${}_N Y_q = {}_N X_I B_p Q_q + {}_N E_q$$

Man sieht dabei, daß diese Form auch eine Verallgemeinerung des multivariaten Modells in der an Querschnittdaten orientierten Varianz- bzw. Regressionsanalyse darstellt, denn der Ansatz der multivariaten Analyse lautet (siehe Timm, 1975) :

$$(5.42) \quad {}_N Y_T = {}_N X_I B_T + {}_N E_T$$

(5.42) ist identisch mit (5.41), wenn  $Q$  die Einheitsmatrix:  ${}_T I_T$  ist.

Da das Wachstumskurvenmodell nur eine spezielle Schreibweise der Varianzanalyse mit abhängigen Messungen darstellt, ist auch klar, daß das Wachstumskurvenmodell keine Neuentwicklung ist, wie Timm schreibt. Es erlaubt aber, in Weiterentwicklung des „gemischten Modells“, die restriktiven Bedingungen für die Störglieder fallen zu lassen.

\* Nach Formel (5.8):  $\text{vec}(ABC) = A \otimes C' \text{vec}(B)$

\*\* Bei allen genannten Autoren werden mehr oder weniger stark unterschiedliche Notationen verwendet. Um nicht noch eine weitere einzuführen, schließen wir uns in der Notation Timm an. Daher muß  $K'$  als  $Q$  geschrieben werden und  $T$  als  $q$ . Die Anzahl der Gruppen  $G$  wird mit  $I$  bezeichnet.

## 5.3 Schätzung des Modells

Für die Ableitung der Schätzer des Modells ist die ausgerollte Form vorteilhafter. Die Störglieder sind nicht unabhängig identisch verteilt. Daher kann keine OLS-Schätzung (Ordinary Least Squares), sondern es muß GLS (Generalized Least Squares) verwendet werden, damit effiziente Schätzer entstehen (siehe Theil, 1970).

Dann ist:\*

$$(5.43) \quad \hat{b} = [(X \otimes K)'(I \otimes \Sigma_e)^{-1}(X \otimes K)]^{-1}[(X \otimes K)'(I \otimes \Sigma_e)^{-1}] y$$

Nach mehreren algebraischen Umformungen mit Hilfe der Regeln für Kroncker-Produkte erhält man:

$$(5.44) \quad \hat{b} = [\{(X'X)^{-1}X'\} \otimes \{(K'\Sigma_e^{-1}K)^{-1}K'\Sigma_e^{-1}\}] y$$

Dieses Ergebnis kann man wieder „einrollen“ (s.a. McDonald & Swaminathan, 1973) :

$$(5.45) \quad \begin{bmatrix} \hat{b}_{11} \dots \hat{b}_{1p} \\ \dots \\ \hat{b}_{l1} \dots \hat{b}_{lp} \end{bmatrix} = \hat{B} = (X'X)^{-1}X'Y\Sigma_e^{-1}K(K'\Sigma_e^{-1}K)^{-1} \\ \text{oder mit } Q=K' \\ = (X'X)^{-1}X'Y\Sigma_e^{-1}Q'(Q\Sigma_e^{-1}Q')^{-1}$$

Bei dieser allgemeinen Schätzung für B ist es notwendig, daß  $\Sigma_e$  bekannt ist. Das kann im allgemeinen nicht vorausgesetzt werden.  $\Sigma_e$  kann aber im Rahmen einer ML-Schätzung (Maximum Likelihood) zusätzlich geschätzt werden. Für einige wichtige Spezialfälle muß die Kenntnis von  $\Sigma_e$  nicht vorausgesetzt werden. Es resultieren daher einfachere Schätzverfahren für folgende Spezialfälle:

1. Bei der multivariaten - an Querschnittsdaten orientierten Varianz- oder Regressionsanalyse ist  $K = I$  (s.o.)

Durch Einsetzen in (5.45) erhält man:

$$\begin{aligned} \hat{B} &= (X'X)^{-1}X'Y\Sigma_e^{-1}I(I'\Sigma_e^{-1}I)^{-1} = \\ &= (X'X)^{-1}X'Y\Sigma_e^{-1}\Sigma_e \quad \text{Da } (A^{-1})^{-1} = A \\ &= (X'X)^{-1}X'Y \end{aligned}$$

Da sich  $\Sigma_e$  wegekürzt, ist die Kenntnis von  $\Sigma_e$  für die Schätzung nicht erforderlich.

\* Während für das Modell:

$y = Zb + e$ ; mit  $\text{Cov}(e) = \sigma^2 V$   
 der OLS-Schätzer  $\hat{b} = (Z'Z)^{-1}Z'y$  ist,  
 hat der GLS-Schätzer die Form:  $\hat{b} = (Z'V^{-1}Z)^{-1}Z'V^{-1}y$   
 hier:  $Z = (X \otimes K)$ ;  $\text{Cov}(e) = I \otimes \Sigma_e = V$

2. Es sei  $K$  eine quadratische Matrix, deren Inverse existiert. Das ist der Fall bei *vollständiger Spezifikation* der Trendparameter.

Falls die Inverse von  $K$  existiert, gilt:

$$(K' \Sigma_e^{-1} K)^{-1} = K^{-1} \Sigma_e K'^{-1} \quad \text{wegen } (AB)^{-1} = B^{-1} A^{-1}$$

Durch Einsetzen in (5.45) erhält man:

$$\begin{aligned} \hat{B} &= (X'X)^{-1} X' Y \underbrace{\Sigma_e^{-1} K (K^{-1} \Sigma_e K'^{-1})}_{\substack{I \\ I}} \\ &= (X'X)^{-1} X' Y K'^{-1} \end{aligned}$$

Falls zusätzlich gilt, daß  $K$  orthonormal ist; gilt damit auch  $K'^{-1} = K$   
 $= (X'X)^{-1} X' Y K$

Diese Lösung ist genau das Produkt der üblichen MANOVA-Schätzung (unter Spezialfall 1.) mit der Kontrastmatrix für die Zeitpunkte:  $K$ .

Für diesen Spezialfall gelten auch die Schätzungen von Bock (1963), Finn (1969) und Bock (1975).

3. Falls die Compound-Symmetry-Annahme für die  $\Sigma_e$ -Matrix zutrifft, kann man zeigen (siehe Morrison (1976), Theil (1971) und Bock (1979)), daß *bei unvollständiger* Spezifikation des Trends gilt:

$$\hat{B} = (X'X)^{-1} X' Y K (K'K)^{-1}; \quad \text{(auch hier ist die Schätzung von } \Sigma_e \text{ nicht notwendig)}$$

Falls  $K'K = I$  (orthonormale Kontraste) erhält man wieder wie unter 2.:

$$\hat{B} = (X'X)^{-1} X' Y K$$

Falls aber eine *unvollständige* Spezifikation des Trends vorliegt ( $K$  ist nicht quadratisch) und keine ganz spezielle Annahme (wie bei Punkt 3.) über  $\Sigma_e$  gemacht werden kann, gelten die einfachen Ergebnisse für den Schätzer von  $B$  nicht mehr. Dann sind auch die Schätzungen, die Bock (1963, 1975) und Finn (1969, S. 408) vorschlagen, nicht effizient.

Gerade dieser Fall ist aber interessant, da man auch Schätzungen erhalten sollte, die nicht (bei gegebenem  $T$ ) auch noch die unwichtigen Polynome höchstmöglicher Ordnung, mitfitten. Ein Grund hierfür könnte die durch einen Adäquatheitstest gewonnene Erkenntnis sein, daß Polynome höherer Ordnung nur insignifikante Beiträge leisten. Finn (1969, S. 408) berichtet einen solchen Fall. Leider schätzt er dann die Parameter fälschlicherweise nach Spezialfall 2.

Khatri (1966) hat einen ML-Schätzer für diesen generellen Fall gefunden.

$$\hat{B} = [(X'X)^{-1} X' Y D^{-1} K (K' D^{-1} K)^{-1}]$$

wobei  $D = {}_T Y'_N (I_N - {}_N X_I (X'_N X_I)^{-1} X'_N) Y$ ;

dabei ist  $D$  der Schätzer von  $(N-k)\Sigma_e$ ;

mit  $k = \text{Rang}(X)$

## 5.4 Hypothesentests

Die Möglichkeit, Hypothesen zu testen, entspricht in diesem allgemeinen zeitbezogenen Modell dem Vorgehen beim multivariaten Querschnittmodell (siehe Timm, 1975):

$$(5.46) \quad \begin{array}{ll} H_0 : {}_g C_l B_p A_u = {}_g \Gamma_u & \text{Nullhypothese} \\ H_1 : C B A \neq \Gamma & \text{Alternativhypothese} \end{array}$$

wobei:

- ${}_g C_l$ : Hypothesenmatrix für Vergleiche zwischen den „Gruppen“ bzw. den Zeilen der Parametermatrix  $B$   
Techn. Anm.: es wird angenommen, daß  $g \leq l$  und  $\text{Rang}(C) = g$
- ${}_p A_u$ : Hypothesenmatrix für Vergleiche „zwischen den Zeiteffekten“, „innerhalb der Gruppen“, bzw. zwischen den Spalten der Parametermatrix  $B$ .  
Techn. Anm.:  $\text{Rang}(A) = u \leq p$
- ${}_g \Gamma_u$ : wird meist 0 gesetzt, es könnten aber auch andere konstante Werte sein.

In allen 3 Matrizen können Werte frei gewählt werden und damit beliebige Hypothesen getestet werden. Die Anzahl der Zeilen von  $C$  und Anzahl Spalten von  $A$  kann im Rahmen der technischen Bedingungen ebenfalls frei gewählt werden. Damit ist die Testung einzelner oder simultaner Mehrfachhypothesen möglich. Die Anzahl der simultan getesteten Hypothesen beläuft sich auf  $g * u$ .

Für den Aufbau der Teststatistiken werden zwei Matrizen mit Fehlerquadratsummen und Fehlerkreuzprodukten berechnet:

$$(5.47a) \quad \begin{array}{ll} \text{a) Hypothesenmatrix} & : Q_H = {}_u A'_p \hat{B}_l' C'_g (C_l' R_l C'_g)^{-1} C_l \hat{B}_p A_u \\ (5.47b) \quad \text{b) Fehlermatrix*} & : Q_E = {}_u A'_p (K' D^{-1} K)^{-1} {}_p A_u \\ & \text{(Errormatrix)} \end{array}$$

wobei außer  $R$  alle Matrizen bereits eingeführt sind; für  $R$  erhält man folgendes, etwas komplizierten Ausdruck (siehe Morrison (1976)):

$$(5.48) \quad R = (X'X)^{-1}(X'X + X'Y[D^{-1} - D^{-1}K(K'D^{-1}K)^{-1}K'D^{-1}]Y'X)(X'X)^{-1}$$

Der komplizierte Ausdruck läßt sich für  $R$  stark vereinfachen, falls  $K$  eine quadratische Matrix ist, für die die Inverse existiert: (bei vollständiger Spezifikation des Trends ist  $p=T$ )

---

\*  $Q_E$  wird manchmal auch als Fehlermatrix des vollen Modells bezeichnet ( $= Q_{\text{voll}}$ )  
 $Q_{\text{restr}} = Q_E + Q_H$  ( $=$  Fehlermatrix des durch die Nullhypothese restringierten Modells)

$$\text{Denn: } (D^{-1} - D^{-1}K(K'D^{-1}K)^{-1}K'D^{-1}) = \\ (D^{-1} - D^{-1}\underbrace{K K^{-1}}_I D \underbrace{K'^{-1}K'}_I D^{-1}) = 0$$

$\underbrace{\hspace{10em}}_I$

$$(5.49) \quad \text{Daher: } R = (X'X)^{-1}$$

Diesem Spezialfall wurde in der Literatur besondere Aufmerksamkeit gewidmet (Bock (1963), Bock (1975), Bock (1979)).

Mit Hilfe von  $Q_H$  und  $Q_E$  können verschiedene Testkriterien aus den Wurzeln der Determinantengleichung  $|Q_H - \Lambda Q_E| = 0$  (Eigenwerte) hergeleitet werden (siehe Morrison):

- Roys - größtes Wurzelkriterium
- Lawley - Hotelling's-Spurstatistik
- Wilks Lambda

Das gebräuchlichste Prüfkriterium ist wohl Wilks Lambda. Es kann als Verhältnis zweier Determinanten errechnet werden:

$$\Lambda = \frac{|Q_E|}{|Q_E + Q_H|}$$

Die Verteilung von  $\Lambda$  ist bekannt als U-Verteilung (siehe Timm (1975)) mit 3 Parametern, die tabelliert vorliegt:

$$\Lambda \sim U(u, g, N - I - (q - p))^*$$

Als Approximation kann auch die  $\chi^2$ -Verteilung genommen werden (nach Bartlett (1947); F-Verteilungs-Approximationen siehe Box (1949), Morrison (1976)).

$-[N - I - (q - p) - \frac{1}{2}(u - g + 1)] \ln(\Lambda)$  ist approximativ  $\chi^2$  verteilt mit  $g \cdot u$  Freiheitsgraden.

*Fortsetzung des Beispiels:*

Angenommen, es sollte die Null-Hypothese  $H_0$  überprüft werden: Es gibt keine Unterschiede im quadratischen und linearen Trend zwischen den Gruppen (keine Überschneidungen der Trendlinien = *keine Interaktion*).

Bei der *gruppenspezifischen* Parametrisierung bedeuten die Koeffizienten der Matrix B:

---

\*  $q$  ist im vorliegenden Fall gleich  $T$ ; bei vollständiger Spezifikation ist  $q - p = 0$ .  $q$  ist bei Multiresponses gleich  $M \times T$

1. Gruppe  
2. Gruppe

$$I \left\{ \begin{array}{ccc} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{array} \right\} = B$$

$\downarrow$  linear Niveau  
 $\downarrow$  quadrat.

Die Hypothese lautet für diese Parametrisierung:

$$b_{12} = b_{22} \Rightarrow b_{12} - b_{22} = 0$$

$$\text{und } b_{13} = b_{23} \Rightarrow b_{13} - b_{23} = 0$$

In der allgemeinen Hypothesenform (5.46) lautet:

$$H_0: \begin{bmatrix} 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = (b_{12} - b_{22}, b_{13} - b_{23}) = (0, 0)$$

$$H_0: {}_1C_2 \cdot B \cdot {}_3A_2 = (0, 0)$$

Bei der *gruppenübergreifenden* Parametrisierung bedeuten die Koeffizienten der Matrix B (siehe oben Seite 359):

Gruppendurchschnitt  
Gruppenunterschied

$$\begin{bmatrix} b_{11}^* & b_{12}^* & b_{13}^* \\ b_{21}^* & b_{22}^* & b_{23}^* \end{bmatrix}$$

$\downarrow$  linear Niveau  
 $\downarrow$  quadrat.

Die Hypothese lautet für diese Parametrisierung:

$$b_{22}^* = 0 \quad \text{und} \quad b_{23}^* = 0$$

In der allgemeinen Hypothesenform (5.46) ist:

$$H_0: \begin{bmatrix} 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} b_{11}^* & b_{12}^* & b_{13}^* \\ b_{21}^* & b_{22}^* & b_{23}^* \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = (b_{22}^*, b_{23}^*) = (0, 0)$$

$$H_0: {}_1C_2 \cdot B \cdot {}_3A_2 = (0, 0)$$

Bei beiden Parametrisierungen resultieren die gleichen Testergebnisse.

### 5.3 Mehrfachantwort (echt multivariate) -Analyse\*

$$T \geq 2; M \geq 2; N \geq T \cdot M$$

Obwohl schon bisher von multivariater Analyse die Rede war (die T-Messungen auf einer Variablen pro Person können als T-Variablen angesehen werden), wurde jeweils nur eine einzige, über die Zeitpunkte hin inhaltlich gleiche Variable betrachtet (z.B.: Entwicklung bezüglich der Schulangst).

Man könnte auch versuchen, simultan mehrere Variablen (z.B.: Schulangst und Lernerfolg) in ihrer Entwicklung über die Zeitpunkte hinweg zu betrachten. Diese Erweiterung ist leicht ins Wachstumskurvenmodell zu integrieren.

Beispiel: Die Variablen „Schulangst“ und „Lernerfolg“ werden bei Schülern und Schülerinnen zu drei Zeitpunkten beobachtet.

Die Modelle für die Analysen der *einzelnen* Variablen wären dann:

Schulangst:

$$\begin{array}{c} 1 \\ \vdots \\ \delta \\ \vdots \\ \varphi \\ \vdots \\ N \end{array} \begin{array}{|c|} \hline E(Y_1 | X) \\ \hline \end{array} = \begin{array}{|c|} \hline X \\ \hline \end{array} \begin{array}{|c|} \hline B_1 \\ \hline \end{array} \begin{array}{|c|} \hline Q_1 \\ \hline \end{array} = X \cdot (B_1 \cdot Q_1)$$

Lernerfolg:

$$\begin{array}{c} 1 \\ \vdots \\ \delta \\ \vdots \\ \varphi \\ \vdots \\ N \end{array} \begin{array}{|c|} \hline E(Y_2 | X) \\ \hline \end{array} = \begin{array}{|c|} \hline X \\ \hline \end{array} \begin{array}{|c|} \hline B_2 \\ \hline \end{array} \begin{array}{|c|} \hline Q_2 \\ \hline \end{array} = X \cdot (B_2 \cdot Q_2)$$

und zusammengefaßt:

$$\begin{array}{c} 1 \dots T \quad 1 \dots T \quad 1 \dots I \quad 1 \dots I \quad 1 \dots T1 \dots T \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \delta \quad \delta \quad \delta \quad \delta \quad \delta \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \varphi \quad \varphi \quad \varphi \quad \varphi \quad \varphi \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ N \quad N \quad N \quad N \quad N \end{array} \begin{array}{|c|c|} \hline E(Y_1 | X) & E(Y_2 | X) \\ \hline \end{array} = \begin{array}{|c|} \hline X \\ \hline \end{array} \begin{array}{|c|c|} \hline B_1 Q_1 & B_2 Q_2 \\ \hline \end{array} = \begin{array}{|c|} \hline X \\ \hline \end{array} \begin{array}{|c|c|} \hline B_1 & B_2 \\ \hline \end{array} \begin{array}{|c|c|} \hline Q_1 & Q_2 \\ \hline \end{array}$$

\* Diese Analyseart heißt bei Bock (1975, S. 502): Analysis of Multiple Repeated Measurement Variables, bei Timm (S. 507): das Multivariate Response Growth Curves Design.



Dabei darf durchaus angenommen werden (unter Umständen nach entsprechenden Tests), daß für die beiden Variablen unterschiedliche Trendkomponenten zutreffen. So kann z.B. für die „Schulungst“ nur ein linearer, für den „Lernerfolg“ sowohl ein linearer wie auch ein quadratischer Trend zutreffen.

Damit hat man als generelles Modell:

$$(5.50) \quad {}_N Y_{(M \cdot T)} = {}_N X_I B_p \underbrace{Q_{M \cdot T}}_q + {}_N \underbrace{E_{M \cdot T}}_q$$

$$\text{wobei } {}_1 B_p = [\underbrace{B_1}_{p_1} | \underbrace{B_2}_{p_2} | \dots | \underbrace{B_M}_{p_M}]; \quad Q = p \left\{ \begin{array}{ccc|c} Q_1 & O & & O \\ O & Q_2 & & O \\ \hdashline & & & \\ O & O & & Q_M \end{array} \right\}$$

und  $E = [E_1 | E_2 | \dots | E_M]$

$\underbrace{\hspace{10em}}_{M \cdot T}$

Das Modell (5.50) weist gegenüber (5.41) eine stärkere Strukturierung der Matrizen auf. So sind die Matrizen  $B$ ,  $Q$  und  $Y$ ,  $E$  aus Teilmatrizen zusammengesetzt. Die Zeilen von  $Y$  sind voneinander unabhängig (Zufallsauswahl der Personen).

Daß auch die Testmöglichkeiten im gleichen Rahmen wie vorher existieren, ergibt sich schon daraus, daß multivariate Tests schon bei  $M = 1$  vorgesehen wurden.

## 6. Pooling von „Querschnitt“- mit „Zeitreihen“-Analyse

Im Jahre 1966 veröffentlichten Balestra, P. & Nerlove, M. einen Artikel, der die Formulierung: „Pooling“ von Querschnitt- und Zeitreihendaten enthielt. Unter diesem Stichwort sind dann in der Folge eine Reihe von Veröffentlichungen entstanden, die sich mit der Konzeption eines solchen Modells und den damit zusammenhängenden Schätzproblemen beschäftigten (z.B. Amemiya, T., 1967; Mundlak, 1978; Hall, 1978; Wallace & Hussain, 1969).

### 6.1 Modellüberlegungen

Ein einfaches Modell in der Zeitreihenanalyse ist das autoregressive 1. Ordnung :

$$(6.1) \quad Y_t = b Y_{t-1} + \varepsilon_t$$

für die  $T$ -Zeitpunkte:

$$\begin{aligned}
 Y_2 &= b Y_1 + \varepsilon_2 \\
 Y_3 &= b Y_2 + \varepsilon_3 \\
 &\vdots \\
 Y_T &= b Y_{T-1} + \varepsilon_T
 \end{aligned}$$

Die  $\varepsilon_t$  (Störglieder) repräsentieren wieder Effekte von Variablen, die nicht explizit im Modell enthalten sind. Es wird angenommen, daß die  $\varepsilon$ 's unkorreliert sind mit Erwartungswert 0 und Varianz  $\sigma_\varepsilon^2$  für alle  $\varepsilon_t$  (Nerlove, 1971, S. 360).

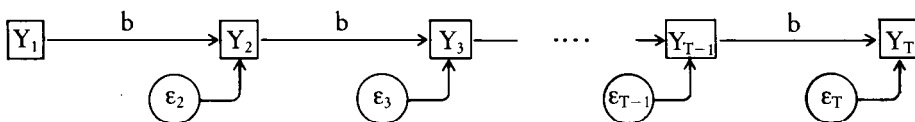


Fig. 6.1: AR(1)-Prozeßmodell

Bei Zeitreihendaten wird für einen Zeitpunkt jeweils nur eine Messung in den Variablen verlangt. Falls aber bei mehreren Personen solche Zeitreihendaten erhoben werden können, sollte pro Person noch ein spezieller Effekt vorgesehen werden, der die Unterschiede zwischen den Personen berücksichtigt: so daß folgendes Modell notwendig wird (bei Annahme einer linearen Wirkung des Personeneffekts:  $\pi_v$ ):

$$(6.2) \quad Y_{vt} = aY_{v,t-1} + \overbrace{\pi_v}^* + \varepsilon_{vt}$$

Als weiteren Schritt, die Variaten in  $Y_t$  adäquat zu beschreiben, schlägt Nerlove (1971) vor, noch zusätzlich einen allgemeinen Zeiteffekt zu berücksichtigen:

$$(6.3) \quad Y_{vt} = aY_{v,t-1} + \pi_v + \tau_t + \varepsilon_{vt}$$

Ein ähnliches Modell ist uns aber bereits aus der Modellgleichung (5.1) bekannt. Allerdings wurde dort  $Y_{v,t-1}$  nicht als unabhängige Variable berücksichtigt. Bei der bedingten Analyse bei 2 Zeitpunkten wurde der Anfangswert als unabhängige Variable mit verwendet. Es war aber kein  $\pi_v$  für die Personeneffekte vorgesehen. Es handelt sich hier also um ein Modell, das im Sinne einer Varianzanalyse ebenfalls Kovariaten (zusätzliche quantitative Variable) enthält.

---

\* In der Formulierung von Nerlove M. (1971, S. 360) wird für:  $\pi_v = \mu_i$  und für  $\varepsilon_{vt} = v_{it}$  gewählt.

Die  $\pi_v$  repräsentieren die Effekte von nicht explizit mit in die Analyse aufgenommenen Variablen, die konstant über die Zeit hinweg wirken.

Das Modell wird meist aber noch allgemeiner angesetzt (siehe z.B.: Mundlak (1978))

$$(6.4) \quad Y_{vt} = \sum_{i=1}^K X_{ivt} b_i + e_{vt} \quad \text{wobei: } e_{vt} = \pi_v + \tau_t + \varepsilon_{vt}$$

in Matrixform: 
$$\begin{bmatrix} y \end{bmatrix} = \begin{bmatrix} X \end{bmatrix} \begin{bmatrix} b \end{bmatrix} + \begin{bmatrix} e \end{bmatrix}$$

Dabei können die  $X$ -Werte einfach wieder verzögerte  $Y$ -Werte (z.B.:  $Y_{v,t-1}$ , ...,  $Y_{v,t-m}$ ) oder zusätzliche unabhängige Variable (kontemporär oder verzögert) sein.

Über die Spezifikation von  $\pi_v$  und  $\tau_t$  ist eine Kontroverse entstanden, die jedem „Varianzanalytiker“ vertraut ist: Sind die Koeffizienten  $\pi_v$  und  $\tau_t$  jeweils fixe Größen (fixer Faktor) oder zufällige Variablen (Zufallsfaktor)?

Im Rahmen des gemischten Modells und der Wachstumskurvenanalyse wurde  $\tau_t$  als fixer Effekt angesehen (mit Reparametrisierungen) und  $\pi_v$  als zufälliger.

Werden beide als fix angesehen, liegt die übliche Spezifikation für das gewöhnliche Regressionsmodell (OLS) vor. Es gibt hier jedoch eine Ausnahme.

Falls zeitverzögerte Variable ( $Y_{t-i}$ ) auf der Seite der unabhängigen Variablen auftauchen, kann man nicht mehr davon ausgehen, daß die unabhängigen Variablen feste Größen im Sinne des klassischen Modells, sondern zufallsbehaftete Variable sind. Die klassischen Schätzmethoden liefern allerdings weiterhin konsistente und effiziente Schätzungen, falls angenommen werden darf, daß die *zufallsbehafteten* unabhängigen Variablen (hier  $Y_t$ ) mit den Störgliedern ( $\varepsilon_t$ ) nicht korrelieren.

Im Rahmen des „Pooling“ wird meist angenommen, daß die Effekte  $\pi_v$  und  $\tau_t$  Zufallsvariable mit den üblichen Annahmen sind (siehe auch oben (S. 344f.)). Es sei die Summe über die Komponenten mit  $e_{vt}$  abgekürzt:

$$(6.5) \quad \begin{array}{ll} \text{a) } e_{vt} = \pi_v + \tau_t + \varepsilon_{vt} & 1. \text{ wobei Kovarianzen zwischen den Komponenten Null sind} \\ \text{b) } \text{Cov}(\pi_v, \tau_t) = \text{Cov}(\pi_v, \varepsilon_{vt}) = \text{Cov}(\tau_t, \varepsilon_{vt}) = 0 & 2. \text{ Kovarianzen zwischen je zwei verschiedenen Variablen einer Komponente sind Null} \\ \text{c) } \begin{array}{l} \text{Var}(\tau_t) = \sigma_\tau^2 \\ \text{Var}(\varepsilon_{vt}) = \sigma_\varepsilon^2 \\ \text{Var}(\pi_v) = \sigma_\pi^2 \end{array} & 3. \text{ Homoskedastizität} \end{array}$$

Als Varianz-Kovarianzmatrix für den Vektor  $e$  ergibt sich damit (mit  $e_v$  als Teilvektor für die  $v$ -te Person):

$$\begin{aligned}
 (6.6) \quad \Sigma_e = \text{cov} \left( \begin{array}{c} T\{e_1 \\ e_2 \\ \dots \\ e_N\} \end{array} \right) &= \begin{array}{|c|c|c|c|} \hline A & B & & B \\ \hline B & A & & B \\ \hline & & & \\ \hline B & B & & A \\ \hline \end{array} = \\
 &= ({}_N I_N \otimes I_T I_T') \sigma_\pi^2 + (I_N I_N' \otimes {}_T I_T) \sigma_\tau^2 + ({}_N I_N \otimes {}_T I_T) \sigma_\epsilon^2 \\
 \text{mit } {}_T A_T &= I_T I_T' \sigma_\pi^2 + {}_T I_T \sigma_\tau^2 + {}_T I_T \sigma_\epsilon^2 = \\
 &= \begin{array}{|c|} \hline \sigma_\pi^2 \dots \sigma_\pi^2 \\ \hline \dots \\ \hline \sigma_\pi^2 \dots \sigma_\pi^2 \\ \hline \end{array} + \begin{array}{|c|} \hline \sigma_\tau^2 \dots 0 \\ \hline \dots \\ \hline 0 \dots \sigma_\tau^2 \\ \hline \end{array} + \begin{array}{|c|} \hline \sigma_\epsilon^2 \dots 0 \\ \hline \dots \\ \hline 0 \dots \sigma_\epsilon^2 \\ \hline \end{array} \\
 {}_T B_T &= {}_T I_T \sigma_\tau^2 = \begin{array}{|c|} \hline \sigma_\tau^2 \dots 0 \\ \hline \dots \\ \hline 0 \dots \sigma_\tau^2 \\ \hline \end{array}
 \end{aligned}$$

Falls die übrigen unabhängigen Variablen als Konstanten aufgefaßt werden können, stimmt die Kovarianzmatrix aller  $y$ -Werte mit der der  $e$ -Werte\* überein.

### Beispiel:

Als abhängige Variable werde Schulangst untersucht. Es kann angenommen werden, daß die Angst zum gegenwärtigen Zeitpunkt ( $Y_t$ ) durch das Ausmaß der Angst im vorherigen Zeitpunkt ( $Y_{t-1}$ ) beeinflusst wird. Andererseits sei das Ausmaß der Angst auch vom empfundenen Strafausmaß ( $X_t$ ) abhängig. Es werden dabei mehrere Personen zu vier Zeitpunkten untersucht.

$\pi_v$  seien die zeitlich konstanten Effekte einer Fülle nicht erfaßter Variablen, die personenspezifisch\*\* auf das Angstniveau Auswirkungen haben.

$\tau_t$  sind die zeitlich wirkenden Effekte nicht explizit erhobener Variablen.

$\epsilon_{vt}$  repräsentieren die personenspezifisch über die Zeit variabel wirkenden Effekte.

Das Modell lautet:

\*5 Dieses Modell haben auch Fuller & Battese (1974) betrachtet und effiziente und konsistente Schätzer dafür entwickelt.

$$Y_{vt} = a_1 Y_{v,t-1} + a_2 X_{vt} + \pi_v + \tau_t + \varepsilon_{vt}$$

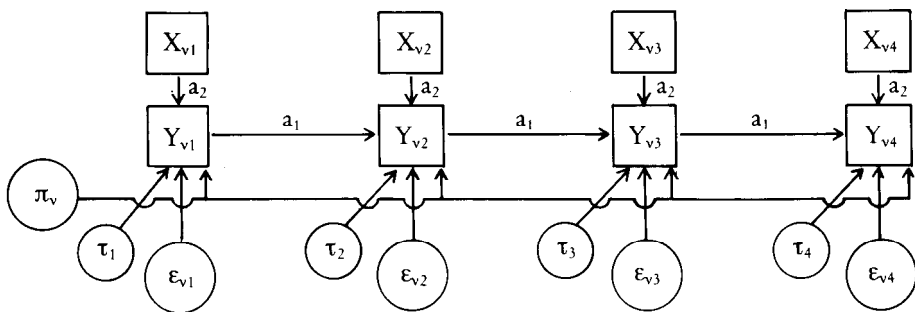


Fig. 6.2: Prozeßmodell für die v-te Person

Als Kritik an einem solchen Modell kann man leicht vorbringen, daß zeitlich weitere Abhängigkeiten zu berücksichtigen wären (siehe oben zur Kritik der Compound-Symmetry):

- a) in den X- bzw. Y-Variablen (diese Abhängigkeiten sind leicht ohne grundsätzliche Änderungen der Form des Modells (6.4) zu berücksichtigen): So könnte man auch unterschiedliche Koeffizienten zwischen Zeitpunkten fordern.
- b) in den  $\tau$ 's oder  $\varepsilon$ 's: (1) Bei den  $\tau$ 's könnte man solche Abhängigkeiten im Sinne des ARIMA-Modells postulieren. (2) Bei den  $\varepsilon$ 's wären solche Abhängigkeiten wieder möglich. Sie würden nicht global sondern spezifisch für jede Person sein. Da Silva (1975) hat ein moving-average-Modell\*\*\* dieser Art und die dazugehörigen Schätzmethoden entwickelt. Dabei setzt er allerdings voraus, daß die gemessenen unabhängigen Variablen nichtstochastisch sind (es dürften in diesem Modell keine zeitverzögerten abhängigen Variablen auf der Prediktorseite auftreten).

Eine interessante Mischung von Abhängigkeiten hat Parks (1967) in sein Modell aufgenommen. Es sieht einen autoregressiven Prozeß 1. Ordnung für die Summe der drei Komponenten

$$e_{vt} = \rho_v e_{v,t-1} + \varepsilon_{vt} \text{ mit } e_{vt} = \pi_v + \tau_t + \varepsilon_{vt}$$

vor. Dabei ist sogar zugelassen, daß die  $\varepsilon_{vt}$  kovariieren.

\*\* Die Personeneffekte sind von vornherein als zeitinvariant konzipiert worden. Daher ist es nicht sinnvoll, bei den  $\pi$ 's Zeitabhängigkeiten anzunehmen. Zeitabhängigkeit tritt allerdings dann ein, wenn  $e_{vt} = \pi_v + \varepsilon_{vt} + \tau_t$  als autoregressiver Prozeß angesetzt wird (Parks, 1967; Jöreskog, 1979).

\*\*\*  $\varepsilon_{vt} = \alpha_0 u_t + \alpha_1 u_{t-1} + \dots + \alpha_M u_{t-M}$  wobei die  $u_t$  unabhängige Zufallsvariable mit gleicher Varianz sind.

## 6.2 Schätzprobleme

Als Schätzer für die Gleichung (6.4) mit der Kovarianzstruktur (6.6) kann der GLS-Schätzer

$$\hat{b}_{\text{GLS}} = (X' \Sigma_e^{-1} X)^{-1} X' \Sigma_e^{-1} y$$

verwendet werden. Da aber  $\Sigma_e$  meist unbekannt ist, muß  $\Sigma_e$  ebenfalls geschätzt werden. Das naheliegendste ist daher, den ML-Schätzer zu nehmen, zumal er im Rahmen von leicht zugänglichen Programmen (z.B. LISREL) implementiert ist. Bei kleinen Stichproben ( $N \leq 50$ ) ist allerdings größte Vorsicht geboten, wie Simulationen gezeigt haben (Nerlove (1971) und Hannan M. T. & Young A. A. (1977)), wenn stochastische unabhängige Variable vorliegen (z.B. ein zeitverzögertes Y als unabhängige Variable). Für kleinere Stichproben sind andere GLS-Schätzverfahren (siehe Nerlove (1971) Fuller & Battese (1974) und Henderson (1971)) vorteilhafter. Die einfachste Strategie *b* zu schätzen, ist allerdings anzunehmen, daß die Komponenten  $\pi_v$ ,  $\tau_t$  fixe Effekte sind, da dann gewöhnliche OLS-Schätzer verwendet werden können mit  $\pi_v$  und  $\tau_t$  als Parameter von „Dummies“. Dies wird auch die „Within-Covarianz“-Technik genannt und wurde auf ihre Anwendbarkeit als Annäherung für die GLS-Techniken untersucht (Wallace & Hussain (1969)). Leider erhält man nur asymptotische (bei  $T \cdot N \rightarrow \infty$ ) Annäherungen, falls zudem keine zeitverzögerten Y als zusätzliche unabhängige Variable auftauchen.

Mundlak (1978) hat in einem Versuch, eine Synthese zwischen den beiden Alternativen „fixe oder zufällige Effekte“ zu finden, nachweisen können, daß der GLS-Ansatz die eventuell bestehende Abhängigkeit der Effekte von der unabhängigen Variablen X vernachlässigt. Falls dieser Einfluß berücksichtigt werden muß, ist der übliche GLS-Schätzer verzerrt, während über die „Within-Covarianz“-Technik der adäquate Schätzer gefunden wird.

## 7. Strukturgleichungsmodelle

$$T \geq 2, M \geq 2, N > \{M, T\}$$

Strukturgleichungsmodelle stellen eine Synthese aus Regression, Pfadanalyse, ökonometrischen Modellen und Faktorenanalyse dar. Mit ihrer Hilfe wird versucht, multivariate Einflußstrukturen zwischen latenten Variablen oder manifesten Indikatoren abzubilden (Jöreskog, 1973, 1979). Im Gegensatz z.B. zu varianzanalytischen Methoden werden explizit Hypothesen über das Geflecht von Variablenbeziehungen aufgestellt und getestet. Dabei werden von den meisten Autoren ganz im Sinne der Pfad- oder Faktorenanalyse meist nur Korrelations- oder Kovarianzmuster der Variablen untersucht (Jöreskog,

1979; Jöreskog & Sörbom, 1976, 1977; Roskam, 1976, 1979). Erst in letzter Zeit wurden auch vereinzelt Arbeiten publiziert, die Mittelwertsstrukturen und Veränderungen auf latenten Variablen untersuchten (Sörbom, 1976, 1979).

Gemäß unserer eingangs erwähnten theoretischen Orientierung befassen wir uns nur am Rande mit rein korrelativen Studien, da hier sowohl Varianz- als Mittelwertsveränderungen unberücksichtigt bleiben und somit keine testbaren zeitbezogenen Hypothesen über Verläufe gemacht werden können. Auch kovarianzorienteerte Modelle wollen wir nur kurz streifen. Wir verweisen hierzu auf die Literatur. Statt dessen wollen wir uns ausführlicher mit der Analyse von strukturierten Mittelwerten auf manifesten und latenten Variablen in Mehrgruppensdesigns befassen. Alle Analysen werden im Rahmen des LISREL-Modells und seiner entsprechenden Terminologie behandelt (s. a. Jöreskog, 1979; Sörbom, 1979).

## 7.1 Kovarianz- und korrelationsorientierte Analysen von Zeitreihen von Querschnitten: Stabilität von Konstrukten

Da in den weiteren Erörterungen immer wieder das LISREL-Modell vorausgesetzt wird, soll es hier kurz behandelt werden. In ein LISREL (linear structural relations model) gehen Zufallsvektoren von latenten abhängigen  $\eta' = (\eta_1, \eta_2, \dots, \eta_m)$  und latenten unabhängigen Variablen  $\xi' = (\xi_1, \xi_2, \dots, \xi_n)$  ein. Zwischen den latenten Variablen besteht die multivariate Regressionsbeziehung

$$(7.1) \quad \begin{array}{c} 1 \\ \vdots \\ m \end{array} \begin{array}{|c|} \hline B \\ \hline \end{array} \cdot \begin{array}{|c|} \hline \eta_1 \\ \eta_2 \\ \vdots \\ \eta_m \\ \hline \end{array} = \begin{array}{c} 1 \\ \vdots \\ m \end{array} \begin{array}{|c|} \hline \Gamma \\ \hline \end{array} \cdot \begin{array}{|c|} \hline \xi_1 \\ \vdots \\ \xi_n \\ \hline \end{array} + \begin{array}{|c|} \hline \xi_1 \\ \xi_2 \\ \vdots \\ \xi_m \\ \hline \end{array} \quad \text{Strukturmodell der latenten Variablen}$$

1 ... m                      1 ... n

mit:  ${}_m B_m$  = Regressionsmatrix mit Gewichten für die latenten endogenen Variablen  $\eta$

${}_m \Gamma_n$  = Regressionsmatrix mit Gewichten für die latenten exogenen Variablen  $\xi$

${}_m \xi_1$  = Vektor mit Gleichungsfehlern

Die Vektoren der latenten  $\eta$  und  $\xi$  können nicht direkt beobachtet werden. Gemessen werden statt dessen ihre Indikatoren  $y' = (Y_1, \dots, Y_p)$  und  $x' =$

$(X_1, \dots, X_q)$ . Die Beziehungen zwischen den Indikatoren und den latenten Variablen werden auch hier wieder durch multivariate Regressionen gekennzeichnet, die Faktorstrukturen der Faktorenanalyse ähneln:

$$(7.2) \quad \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} = \begin{bmatrix} \Lambda_y \end{bmatrix} \cdot \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \quad \begin{bmatrix} X_1 \\ \vdots \\ X_q \end{bmatrix} = \begin{bmatrix} \Lambda_x \end{bmatrix} \cdot \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_q \end{bmatrix}$$

mit:  $\Lambda_y$  = Regressionsmatrix der  $y$  auf die  $\eta$   
 $\Lambda_x$  = Regressionsmatrix der  $x$  auf die  $\xi$   
 $\varepsilon, \delta$  = Meßfehlervektoren

In dem einfachen LISREL-Modell werden eine Reihe von Annahmen eingebracht, die nachher z.T. fallengelassen werden können. So wird angenommen, daß  $E(\eta) = E(\xi) = E(\zeta) = E(\delta) = E(\varepsilon) = 0$  und  $E(\eta \cdot \varepsilon') = E(\xi \cdot \delta') = E(\xi \cdot \zeta') = 0$  und  $B$  nichtsingulär ist. Die Kovarianzmatrix der Indikatoren ist dann

$$(7.3) \quad \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} = \begin{bmatrix} \Lambda_y (B^{-1} \Gamma \Phi \Gamma' B'^{-1} + B^{-1} \Psi B'^{-1}) \Lambda_y' + \Theta_\varepsilon & \Lambda_y B^{-1} \Gamma \Phi \Lambda_x' \\ \Lambda_x \Phi \Gamma' B'^{-1} \Lambda_y' & \Lambda_x \Phi \Lambda_x' + \Theta_\delta \end{bmatrix}$$

mit folgenden Kovarianzmatrizen der latenten Variablen:

$$(7.4) \quad \begin{array}{c} \begin{array}{c} \varepsilon \\ \delta \\ \zeta \\ \eta \\ \xi \end{array} \end{array} \begin{array}{ccccc} \begin{array}{c} \varepsilon' \\ \delta' \\ \zeta' \\ \eta' \\ \xi' \end{array} & \begin{array}{c} \Theta_\varepsilon \\ 0 \\ 0 \\ 0^* \\ 0 \end{array} & \begin{array}{c} 0 \\ \Theta_\delta \\ 0 \\ 0 \\ 0^* \end{array} & \begin{array}{c} 0 \\ 0 \\ \Psi \\ 0 \\ 0^* \end{array} & \begin{array}{c} 0^* \\ 0 \\ 0 \\ B^{-1} \Gamma \Phi \Gamma' B'^{-1} + B^{-1} \Psi B'^{-1} \\ \Phi \Gamma' B'^{-1} \end{array} & \begin{array}{c} 0 \\ 0^* \\ 0^* \\ B^{-1} \Gamma \Phi \\ \Phi \end{array} \end{array}$$



Die dick umrandeten Matrizen enthalten Parameter, die direkt zu schätzen sind. Dabei gibt es neben unbekannten freien Parametern solche, die einander gleich sein sollen (constrained Parameters) oder andere, die auf einen bestimmten Wert fixiert sind (mit \* markiert). Die P-Matrizen in (7.4) sind nach Voraussetzung Null. Nicht markierte Kästchen in (7.4) enthalten ebenfalls verschwindende Kovarianzen. Die Parameter in den umrandeten Matrizen werden mit Maximum-Likelihood-Methoden geschätzt. Weitere Einzelheiten (auch über das Identifikationsproblem und über die Likelihoodquotiententests zum Hypothesentesten etc.) finden sich bei Jöreskog (1973) und Long (1976).

Wir wollen die Methode an einem Test-Retest-Design mit 2 Variablen zu jedem Meßzeitpunkt demonstrieren. Die beiden Variablen sollen Indikatoren für ein Konstrukt sein. Ist man daran interessiert, ob das Konstrukt im *korrelativen* Sinn über die Zeit hinweg konstant geblieben ist, muß man  $\varrho(\xi, \eta)$  schätzen und gegen die Hypothese  $\varrho(\xi, \eta) = 1.0$  testen. Zu  $t_0$  werden 2 Indikatoren  $X_1$  und  $X_2$  der latenten Variablen  $\xi$  gemessen. Zum zweiten Zeitpunkt  $t_1$  werden die beiden Indikatoren  $Y_1$  und  $Y_2$  beobachtet. Das zugrunde liegende Pfadmodell ist in Figur 7.1 und (7.5) sowie (7.6) dargestellt.

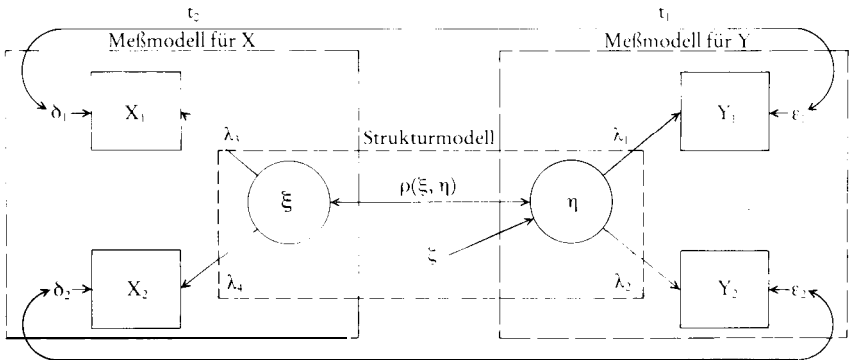


Fig. 7.1: Pfadmodell „Zwei-Wellen-Zwei-Variablen“ zur Überprüfung der korrelativen Stabilität des Konstrukts  $\xi$

$$(7.5a) \quad \begin{bmatrix} Y_1 \\ Y_2 \\ X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0^* \\ \lambda_2 & 0^* \\ 0^* & \lambda_3 \\ 0^* & \lambda_4 \end{bmatrix} \begin{bmatrix} \eta \\ \xi \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \delta_1 \\ \delta_2 \end{bmatrix} \quad \theta_\varepsilon = \begin{bmatrix} \sigma_{\varepsilon_1}^2 & \cdot & \cdot & \cdot \\ 0^* & \sigma_{\varepsilon_2}^2 & \cdot & \cdot \\ \sigma_{\varepsilon_1 \delta_1} & 0^* & \sigma_{\delta_1}^2 & \cdot \\ 0^* & \sigma_{\varepsilon_2 \delta_2} & 0^* & \sigma_{\delta_2}^2 \end{bmatrix}$$

$$(7.5b) \quad y = A_y \eta + \varepsilon \quad \text{alle } Y, X \text{ als Abweichungswerte}$$

Wir gehen gleich von korrelierten Fehlervariablen  $E(\delta_1, \varepsilon_1) \neq 0$  und  $E(\delta_2, \varepsilon_2) \neq 0$  aus. Wir vermuten also, daß die Korrelation zwischen den Variablen  $X_1$  und

$Y_1$  sowie  $X_2$  und  $Y_2$  höher ist als es nur auf Grund ihrer Abhängigkeit von  $\xi$ ,  $\eta$  und deren Korrelation  $\varrho(\xi, \eta)$  zu vermuten wäre. In dem LISREL-Modell (7.5b) fällt uns auf, daß durch Umbenennung  $\eta_1 \leftarrow \eta$ ,  $\eta_2 \leftarrow \xi$ ,  $Y_3 \leftarrow X_1$ ,  $y_4 \leftarrow X_2$ ,  $\varepsilon_3 \leftarrow \delta_1$ ,  $\varepsilon_4 \leftarrow \delta_2$  das Meßmodell (7.2)  $x = A_x \xi + \delta$  nicht mehr auftaucht („no X Option“). Dieses zunächst verwirrende Vorgehen ist aber notwendig, wie ein Blick auf (7.4) zeigt, da im „normalen“ LISREL-Modell (7.1-7.4) Kovarianzen zwischen  $\delta, \varepsilon$  nicht vorgesehen sind. Im Strukturmodell (7.6)

$$(7.6a) \quad \begin{bmatrix} 1^* & 0^* \\ 0^* & 1^* \end{bmatrix} \begin{bmatrix} \eta \\ \xi \end{bmatrix} = \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \quad \Psi = \begin{bmatrix} 1^* & \cdot & \cdot \\ \rho(\xi, \eta) & 1^* & \cdot \\ \cdot & \cdot & 1^* \end{bmatrix}$$

$$(7.6b) \quad B \cdot \eta = \zeta \quad * \text{ fixierter Wert}$$

werden dann durch die Gleichsetzung  $B = Z$  die ursprünglichen „Gleichungsfehler“  $\zeta$  umdefiniert zu  $\zeta_1 \Rightarrow \eta$  und  $\zeta_2 \Rightarrow \xi$ . Das ist sinnvoll, weil die Kovarianzmatrix  $\Psi$  der Gleichungsfehler (hier jetzt: der latenten Variablen  $\xi, \eta$ ) direkt als Parameter schätzbar ist (vgl. 7.4). Damit statt der  $\text{Kov}(\xi, \eta)$  die  $\text{Korr}(\xi, \eta)$  geschätzt wird, legen wir  $\psi_{11} = \psi_{22} = 1.0$  fest.

Es stehen den 10 Kovarianzen der Indikatoren 11 unbekannte Parameter ( $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_2}^2, \sigma_{\delta_1}^2, \sigma_{\delta_2}^2, \sigma_{\delta_1 \varepsilon_1}, \sigma_{\delta_2 \varepsilon_2}$  und  $\varrho(\xi, \eta)$ ) gegenüber. Das Modell ist nicht identifiziert. Eine Vereinfachung würde es bedeuten, wenn man von parallelen Prätests  $X_1$  und  $X_2$  ausgehen könnte. In diesem Falle könnte man  $\lambda_3 = \lambda_4$  setzen. Das Modell wäre dann gerade identifiziert: alle Parameter sind eindeutig aus Kovarianzen und Varianzen schätzbar. Wir wollen dieses Modell als „volles Modell“ bezeichnen.

Da das Meßmodell  $x = A_x \xi + \delta$  fehlt, vereinfacht sich die Kovarianzmatrix (7.3) zu (7.7)

$$(7.7) \quad \Sigma = A_y (B^{-1} \Psi B'^{-1}) A_y' + \Theta_\varepsilon$$

und da  $B = Z$  gesetzt wurde zu (7.8) auf S. 380.

Die Korrelation zwischen den latenten Variablen ist dann:

$$(7.9) \quad \varrho(\xi, \eta) = \left[ \frac{\text{Kov}(X_2, Y_1) \cdot \text{Kov}(X_1, Y_2)}{\text{Kov}(X_1, X_2) \cdot \text{Kov}(Y_1, Y_2)} \right]^{1/2} = \left[ \frac{\lambda_1 \lambda_4 \varrho \cdot \lambda_2 \lambda_3 \varrho}{\lambda_4^2 \cdot \lambda_1 \lambda_2} \right]^{1/2}$$

Sie kann durch die Stichprobenkovarianzen der Kovarianzmatrix  $S$  geschätzt werden.

Um die Hypothese  $\varrho(\xi, \eta) = 1.0$  zu testen, wird das „volle Modell“ durch die Restriktion  $\psi_{21} = 1.0$  in das „reduzierte Modell“ überführt. Das reduzierte Modell ist dann überidentifiziert, da es nur noch 9 Parameter besitzt.

$$(7.8a) \quad \Sigma = A_y \Psi A_y' + \theta_e$$

$$(7.8b) \quad \Sigma = \begin{bmatrix} \text{Var}(Y_1) & \cdot & \cdot & \cdot \\ \text{Kov}(Y_2, Y_1) & \text{Var}(Y_2) & \cdot & \cdot \\ \text{Kov}(X_1, Y_1) & \text{Kov}(X_1, Y_2) & \text{Var}(X_1) & \cdot \\ \text{Kov}(X_2, Y_1) & \text{Kov}(X_2, Y_2) & \text{Kov}(X_2, X_1) & \text{Var}(X_2) \end{bmatrix}$$

$$\begin{aligned} \text{Matrixmultiplikation} &= \begin{bmatrix} \lambda_1 & 0^* \\ \lambda_2 & 0^* \\ 0^* & \lambda_4 \\ 0^* & \lambda_4 \end{bmatrix} \begin{bmatrix} 1^* & \rho(\xi, \eta) \\ \rho(\xi, \eta) & 1^* \end{bmatrix} \begin{bmatrix} \lambda_1 & \lambda_2 & 0^* & 0^* \\ 0^* & 0^* & \lambda_4 & \lambda_4 \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & \lambda_1 \rho & \lambda_1^2 & \cdot & \cdot & \cdot \\ \lambda_2 & \lambda_2 \rho & \lambda_1 \lambda_2 & \lambda_2^2 & \cdot & \cdot \\ 0^* & \lambda_4 & \lambda_1 \lambda_4 \rho & \lambda_2 \lambda_4 \rho & \lambda_4^2 & \cdot \\ 0^* & \lambda_4 & \lambda_1 \lambda_4 \rho & \lambda_2 \lambda_4 \rho & \lambda_4^2 & \lambda_4^2 \end{bmatrix} + \begin{bmatrix} \sigma_{\epsilon_1}^2 & \cdot & \cdot & \cdot \\ 0^* & \sigma_{\epsilon_2}^2 & \cdot & \cdot \\ \sigma_{\delta_1 \epsilon_1} & 0^* & \sigma_{\delta_1}^2 & \cdot \\ 0^* & \sigma_{\delta_2 \epsilon_2} & 0^* & \sigma_{\delta_2}^2 \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1^2 + \sigma_{\epsilon_1}^2 & \cdot & \cdot & \cdot \\ \lambda_1 \lambda_2 & \lambda_2^2 + \sigma_{\epsilon_2}^2 & \cdot & \cdot \\ \lambda_1 \lambda_4 \rho + \sigma_{\delta_1 \epsilon_1} & \lambda_2 \lambda_4 \rho & \lambda_4^2 + \sigma_{\delta_1}^2 & \cdot \\ \lambda_1 \lambda_4 \rho & \lambda_2 \lambda_4 \rho + \sigma_{\delta_2 \epsilon_2} & \lambda_4^2 & \lambda_4^2 + \sigma_{\delta_2}^2 \end{bmatrix} \end{aligned}$$

Maximum-Likelihoodschätzungen des Parametervektors  $\Theta$  erhält man bei multivariater Normalverteilung der Indikatoren, wenn man die Anpassungsfunktion

$$F \{ \Sigma(\Theta) \} = \ln |\Sigma(\Theta)| + \text{Spur} \{ \Sigma^{-1}(\Theta) \} - \ln |S| - p$$

(mit  $p$  = Zahl der Indikatoren) minimiert. Ist die aus den Modellparametern  $\Theta$  rückgerechnete Kovarianzmatrix  $\Sigma(\Theta)$  gleich der Stichprobenkovarianzmatrix  $S$ , ist  $F = 0$ . Sollen zusätzlich zu den Kovarianzen z.B. im Rahmen der multivariaten Wachstumskurvenanalyse auch noch der Mittelwertsvektor  $\mu_z(\Theta)$  an den Stichprobenmittelwertsvektor  $\bar{z}$  angepaßt werden, muß

$$F \{ \Sigma(\Theta), \mu_z(\Theta) \} = \ln |\Sigma(\Theta)| + \text{Spur} \{ \Sigma^{-1}(\Theta) \} + (\bar{z} - \mu_z(\Theta))' \Sigma^{-1}(\Theta) (\bar{z} - \mu_z(\Theta)) - \ln |S| - p$$

minimiert werden. Die Variable  $(N-1) \cdot F$  verteilt sich nach der Chi-Quadratverteilung.

Hypothesentests erfolgen nach der Likelihoodquotientenmethode. Es können nur ineinander „geschachtelte“ Modelle getestet werden. Das reduzierte Mo-

dell (hier mit  $p_{\text{red}} = 9$  Parametern) wird zugunsten des nicht eingeschränkten, vollen Modells (hier mit  $p_{\text{voll}} = 10$  Parametern) verworfen, wenn

$$\chi^2_{\text{red}} - \chi^2_{\text{voll}} = N \cdot F_{\text{red}} - N \cdot F_{\text{voll}} > \chi^2_{1-\alpha, df_{\text{diff}}} \quad (\text{hier: } \chi^2_{.95, df=1} = 3.84)$$

wobei:  $F_{\text{voll}}$  in diesem Fall gleich Null ist

$$df_{\text{diff}} = df_{\text{red}} - df_{\text{voll}} \quad (\text{hier: } 1 - 0 = 1)$$

$$df_{\text{red}} = \text{Anzahl der unabhängigen Parameter } p_{\text{voll}} - \text{Anzahl der unabhängigen Parameter } p_{\text{red}} \quad (\text{hier } 10 - 9 = 1)$$

$$df_{\text{voll}} = \text{Anzahl der Parameter } p_{\text{voll}} - \text{Anzahl der Parameter, die man benötigt, um } S \text{ bzw. } \bar{Z} \text{ perfekt anzupassen (hier: 0)}$$

Die Stabilitätsschätzung der Konstrukte (7.9) ist trotz ihrer methodischen Eleganz nur von geringem Wert. Läßt sich die Nullhypothese  $\varrho(\xi, \eta) = 1.0$  nicht zurückweisen, können dennoch Mittelwerts- und Varianzveränderungen eingetreten sein. Andererseits bedeutet die Verwerfung der Nullhypothese *nicht*, daß die Konstrukte verschieden sind. Es kann sein, daß einige Personen äußeren Einflüssen unterlagen, die ihre Position auf den Indikatoren relativ zu den anderen Personen verändert haben. So wäre es bei einer Stichprobe von  $N$  Personen denkbar, daß eine Teilstichprobe zwischen Prä- und Posttest (Gewichtsmessungen auf 2 Waagen, die als Indikatoren dienen) eine individuell angepaßte Gewichtsreduktion mitmacht. Eine andere Stichprobe nimmt an einer Kreuzfahrt auf einem Luxusdampfer (durchschnittliche Gewichtszunahme ca. 2 kg pro Woche) teil. Der Rest achtet penibel auf sein Gewicht. Die Korrelation des Konstrukts  $\xi$  (= wahres Gewicht, gemessen auf 2 Waagen  $X_1, X_2$  zum Zeitpunkt  $t_0$ ) mit dem Konstrukt  $\eta$  (= wahres Gewicht gemessen auf 2 anderen Waagen  $Y_1, Y_2$  zum Zeitpunkt  $t_1$ ) dürfte nahe Null liegen. Der Schluß,  $\xi$  und  $\eta$  wären daher inhaltlich verschieden oder die Meßinstrumente unbrauchbar, ist falsch, da man die substantiellen exogenen Einflüsse nicht explizit mitberücksichtigt. Solche exogenen Einflüsse sind dabei die Intensität der Kur, des Trainings etc.

Modelle, die die endogene Entwicklung und exogene Einflüsse berücksichtigen, sind alle den systemtheoretischen Modellen zuzurechnen. Diese werden wir in Kapitel 7.3 als zeitdiskrete und in Kapitel 9. als zeitkontinuierliche behandeln.

## 7.2 Wachstumskurvenanalyse als Strukturgleichungsmodell

Wie wir in Kapitel 5. und 3.2 gesehen haben, gestalten sich Schätzung und Hypothesenprüfung im multivariaten Wachstumskurvenmodell (3.2.7) bzw. (5.41) und im echt multivariaten Modell (Kap. 3.2.3) bzw. (5.50) als besonders schwierig, wenn im Modell



und das Meßmodell für Gruppe  $g = 1, 2, \dots, G$  für  $Y$ :

$$y = A_y \cdot \eta + 0$$

$$= \begin{bmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & 1 & & & & & & \\ & & & 1 & & & & & \\ & & & & 1 & & & & \\ & & & & & 1 & & & \\ & & & & & & 1 & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 3 & 9 \end{bmatrix} \cdot \begin{bmatrix} e_{g_1} \\ e_{g_2} \\ e_{g_3} \\ e_{g_4} \\ b_{g_5} \\ b_{g_1} \\ b_{g_2} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

konst.      lin.      quadr.

und für  $X$ :  $X = \xi = 1$

Durch Verlängerung der Vektoren  $y, \eta, I, \xi$  und Erweiterung der Matrizen  $A_y, B$  entlang der Hauptdiagonalen erhält man die LISREL-Spezifikation für die *echt* multivariate Wachstumskurvenanalyse für  $Y_1, Y_2, \dots, Y_m$ .

Sind die Fehler  $e_{g1}, \dots, e_{gT}$  autokorreliert, muß die Submatrix in der linken oberen Ecke von  $B$  so spezifiziert werden, daß  $T \cdot e = \varepsilon$  ist mit unabhängigen  $\varepsilon$ . Für einen autoregressiven Prozeß 1. Ordnung  $e_t = \beta_t e_{t-1} + \varepsilon_t$  nimmt die Matrix

$$B = \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} t_1 & 1 & & & & & 0 \\ t_2 & -\beta_2 & 1 & & & & 0 \\ t_3 & & -\beta_3 & 1 & & & 0 \\ t_4 & & & -\beta_4 & 1 & & 0 \\ & 0 & & & & 1 & 1 \\ & & & & & & 1 & 1 \end{bmatrix}$$

Es lassen sich dann die in LISREL möglichen Hypothesen (Prüfung der Gleichheit und des Verschwindens von Parametern innerhalb und zwischen den Gruppen) prüfen.

Will man z.B. einen *Schereneffekt* der Mittelwertsverläufe (*Interaktion zwischen Gruppen und Zeit*) prüfen, stellt man folgende Nullhypothese auf:

$$H_0: \text{keine Interaktion } b_{11} = b_{21} \text{ (Gleichheit des lin. Trends)} \\ b_{12} = b_{22} \text{ (Gleichheit des quadr. Trends)}$$

Der Haupteffekt der Zeit wird geprüft mit folgender Nullhypothese:

$$\begin{aligned} H_0: \text{kein Zeiteffekt } b_{11} = b_{21} = 0 & \text{ (Verschwinden d. lin. Trends)} \\ b_{12} = b_{22} = 0 & \text{ (Verschwinden d. quadr. Trends)} \end{aligned}$$

Für den Haupteffekt der Gruppen benötigt man die Nullhypothese:

$$\begin{aligned} H_0: \text{kein Gruppeneffekt } b_{10} = b_{20} & \text{ (identische Mittelwerts-} \\ b_{11} = b_{21} & \text{ Verläufe)} \\ b_{12} = b_{22} & \end{aligned}$$

### 7.3 Erwartungswertorientierte Analysen von Zeitreihen von Querschnitten: Zeitbezogene Hypothesen für diskrete Zeitpunkte

Bei der Evaluation von Interventionsprogrammen in Kontrollgruppendesigns treten oft nichtsignifikante Ergebnisse auf. Es gibt nun verschiedene inhaltliche Erklärungen für solche Ergebnisse, die mit dem Prozeßcharakter der Untersuchung zusammenhängen. So kann z.B. ein *Dosierungsproblem* (Möbus, 1981) vorliegen: die Intervention war *effektiv* aber *zeitlich* zu *kurz* angesetzt, um statistisch signifikant zu sein. Es kann auch sein, daß die gesamte Untersuchung wegen des u. U. langsamen Ablaufs der psychischen Lernvorgänge zeitlich *zu komprimiert* war. Die gesamte Untersuchung müßte eventuell verlängert werden, um signifikante Effekte zu sichern. So kann sich die Frage stellen: „Von welchem Zeitpunkt an erwarte ich einen signifikanten Unterschied in den Mittelwertsvektoren zwischen Kontroll- und Experimentalgruppe. Diese Frage kann mit klassischen statistischen Verfahren (Varianz-, Faktorenanalyse etc.) nicht geleistet werden, weil sie als statische Modelle die zeitliche Prozeßstruktur der empirischen Phänomene nur ungenügend modellieren. Selbst die für die Analyse zeitbezogener Daten gerne herangezogene Varianzanalysevariante „Wachstumskurvenanalyse“ (Khatrı, 1966; Timm, 1975) besteht im Prinzip nur aus einer Polynomannpassung der Mittelwertsverläufe. Eine inhaltliche Begründung für die Beschreibung von Entwicklungsverläufen mit Polynomen wird in der Regel nicht gegeben.

Statt dessen wollen wir Längsschnittdaten in diskreter Zeit mit Differenzengleichungssystemen und in kontinuierlicher Zeit mit Differentialgleichungssystemen (Kap. 9) analysieren. Beide Modelltypen sind im Gegensatz zu varianzanalytischen Methoden in dem Sinne dynamisch, daß alle Variablen bzw. Variablenänderungen explizit Funktionen der Zeit sind. Mit den dynamischen Modellen läßt sich die *Transferstruktur* der Variablen untereinander und der Einfluß exogener Variabler abschätzen. Anschließend sind mit den so gewon-

nenen Parametern *Prognosen* und *Simulationen* der erwarteten Entwicklungsverläufe möglich (u.a. Markus & Zajonc, 1977). Im Gegensatz zu den kovarianzorientierten „expost“-Analysen ist dabei die *Formulierung zeitbezogener Hypothesen* möglich: „Wir erwarten unter bestimmten Rahmenbedingungen zu einem *vorher festgelegten* Zeitpunkt einen Mittelwertsvektor  $\mu(t)$  mit den Werten  $\mu_1(t) \dots \mu_m(t)$ .“ Diese Hypothese läßt sich dann multivariat (z.B. mit Hotelling's  $T^2$ ) testen. Dabei ist zu erwarten, daß wir „gegen uns arbeiten“: je größer die Prognosedistanz, desto größer die Wahrscheinlichkeit der Modellfalsifikation. Muß die Prognose falsifiziert werden, gibt es dafür u.a. fünf Erklärungen: (a) vom Forscher nicht bemerkte Änderung des in der Empirie „geltenden“ Gesetzes, (b) Formulierung eines von Anfang an falschen Modells für das Prozeßphänomen, (c) Nichtbeachtung relevanter Variabler, (d) mangelnde Kontrolle der exogenen Randbedingungen oder deren falsche Vorhersage, (e) falsche Messung des Anfangszustandes (Pretest). Die Testung eines Modells entlang der Zeitachse verläuft offensichtlich wesentlich strenger als Modellkontrollen oder Anpassungstests in einem nicht zeitlich aufgespannten Rahmen. Zur Frage der Prognosedistanz sei auf Drösler (1976) verwiesen.

Nehmen wir z.B. an, daß die Daten eines Test-Retest-Designs Realisationen eines multivariaten autoregressiven Prozesses 1. Ordnung „in diskreter“ Zeit sind. Kompliziertere Modelle (s. Chow, 1975) können wegen der Beschränkung auf 2 Meßzeitpunkte nicht formuliert werden. Die Prozeßgleichung lautet:

$$(7.11a) \quad E \{ X(t) | X(t-1) \} = A X(t-1) + b$$

oder für die Mittelwerte

$$(7.11b) \quad \mu(t) = A \mu(t-1) + b$$

Die Parametermatrix  $A$  spiegelt die Einflüsse der Pretests auf die Posttests wider. Sie stellt also die zeitdiskrete Transferstruktur dar. Dagegen repräsentiert der Parametervektor  $b$  die Stärken äußerer Einflüsse, die zu einer Scheinvariablen mit dem Wert 1 für alle Variablen und Zeitpunkte zusammengefaßt wurden, auf die Posttestvariablen  $X(t)$  wieder (s.a. Steyer, 1980).  $A$  und  $b$  können mit LISREL als multivariate Regression der Posttests auf die Pretests geschätzt werden:

$$(7.12a)$$

$$\begin{array}{|c|} \hline 1^* & 0^* \\ \vdots & \vdots \\ 0^* & 1^* \\ \hline \end{array} \cdot \begin{array}{|c|} \hline X_1(t) = \eta_1 \\ \vdots \\ X_m(t) = \eta_m \\ \hline \end{array} = \begin{array}{|c|c|} \hline & \\ \hline A & b \\ \hline \end{array} \begin{array}{|c|} \hline X_1(t-1) = \xi_1 \\ \vdots \\ X_m(t-1) = \xi_m \\ \hline 1 = \xi_{m+1} \\ \hline \end{array} + \begin{array}{|c|} \hline \xi_1 \\ \vdots \\ \xi_m \\ \hline \end{array}$$



$$(7.12b) \quad B \eta = \Gamma \xi + \zeta \quad A_y = I \quad A_x = I \quad \Theta_e = 0 \quad \Theta_\delta = 0$$

Zur Simulation der erwarteten Entwicklungsverläufe werden die Lösungen des Differenzengleichungssystems

$$(7.13a) \quad E\{X(t) | X(0)\} = A^t X(0) + \left\{ \sum_{i=0}^{t-1} A^i \right\} b = \left\{ \begin{array}{c} \text{endogene} \\ \text{Entwicklung} \\ \text{seit } t=0 \end{array} \right\} + \left\{ \begin{array}{c} \text{auf exogenen} \\ \text{Einfluß rück-} \\ \text{führbare} \\ \text{Entwicklung} \end{array} \right\}$$

oder für die Mittelwerte

$$(7.13b) \quad \mu(t) = A^t \mu(0) + \left\{ \sum_{i=0}^{t-1} A^i \right\} b = R A^t R^{-1} \mu(0) + \left\{ \sum_{i=0}^{t-1} R A^i R^{-1} \right\} b$$

berechnet, wobei

- $A^t$  = Transitionsmatrix des Anfangszustandes
- $\mu(t)$  = Mittelwertsvektor zum Zeitpunkt  $t$
- $\mu(0)$  = Mittelwertsvektor zum Zeitpunkt 0 (= Anfangszustand)
- $\left\{ \sum A^i \right\}$  = Transitionsmatrix des konstanten Inputs  $b \cdot 1$
- $A^i$  =  $i$ -faches Produkt von  $A$  (z.B.:  $A^2 = AA$ )
- $R$  = Matrix mit rechten Eigenvektoren von  $A$
- $\Lambda$  = Diagonalmatrix mit Eigenwerten von  $A$  ( $A$  muß diagonalisierbar sein)

Das Verhalten eines *dynamischen* Systems (7.11; 7.13) ist am leichtesten in der Basis der Eigenvektoren (s. 7.13b, rechts) beschreibbar. Hier ist das System entkoppelt, d.h. die Variablen beeinflussen sich wechselseitig nicht mehr. Wachstum, Schrumpfung oder Oszillation der neuen entkoppelten Variablen werden durch die Eigenwerte beschreibbar. Ist ein Eigenwert vom Betrage größer als 1, ist das System instabil: die Variablen wachsen ohne Grenze. Sind alle Eigenwerte von  $A$  vom Betrage kleiner als 1, ist das System asymptotisch stabil: es strebt einer Ruhelage zu, wenn sich die äußeren Einflüsse nicht ändern. Sind die Eigenwerte konjugiert komplex, treten Oszillationen in den erwarteten Mittelwertsverläufen auf.

Es liegt hier ein Vergleich mit der Hauptkomponentenanalyse nahe. Dort sind in der Basis der Eigenvektoren nicht dynamische sondern *statische* Verhältnisse entkoppelt. Die neuen Variablen (= Hauptkomponenten) korrelieren nicht mehr. Die Eigenwerte reflektieren nicht die dynamische sondern die statische Wichtigkeit der neuen entkoppelten Variablen im Sinne von Varianzaufklärungen. Jedoch treten in der Hauptkomponentenanalyse keine konjugiert komplexen oder negativen Eigenwerte auf, da dort statt einer nichtsymmetrischen Regressionsmatrix eine symmetrische positiv semidefinite(gramsche) Korrela-

tionsmatrix faktorisiert wird. Statische Verhältnisse zwischen Variablen werden als symmetrische Korrelationsbeziehungen, dynamische Verhältnisse als nichtsymmetrische Regressionsbeziehungen aufgefaßt.

Man kann die Zusammenfassung aller exogener Einflüsse in *eine* Scheinvariable fallen lassen und statt dessen  $n$  exogene Einflüsse betrachten. Dann erweitert sich (7.11) zu (7.14) und (7.13) zu (7.15):

$$(7.14) \quad \mu(t) = A\mu(t-1) + B\mu(t-1) \quad {}_mB_n = \text{Matrix mit Gewichten} \\ \text{der exogenen} \\ \text{Variablen}$$

$$(7.15) \quad \mu(t) = A^t\mu(0) + \sum_{i=0}^{t-1} A^i B\mu(t-1-i)$$

Damit ist (7.14) der zeitdiskrete Spezialfall von (9.6) und (7.15) von (9.3)). Wenn man gewillt ist, nur Prognosen für regelmäßig wiederkehrende Zeitpunkte zu treffen und zu akzeptieren, daß zwischen diesen Zeitpunkten Unkenntnis über die Zustände des multivariaten Systems herrscht, kann man sich mit den einfacheren zeitdiskreten Modellen begnügen. Sollen jedoch die Prozeßparameter im Sinne von psychologischen Theorien „kausal“ interpretiert werden, ist die Wahl eines zeitdiskreten oder zeitkontinuierlichen Modells für die Interpretationen entscheidend.

Ist der Parameter  $a_{ij}$  im zeitkontinuierlichen Modell gleich Null, ist der entsprechende Parameter im zeitdiskreten Modell (7.15) sehr wahrscheinlich ungleich Null! Behauptet man also im zeitkontinuierlichen Modell, daß die Variable  $i$  den Zuwachs oder Veränderung von Variable  $j$  nicht steuert, muß man im zeitdiskreten Modell das Gegenteil behaupten:  $i$  beeinflusst  $j$ !

Als Beispiel für die Anwendung eines zeitdiskreten Modells soll die Auswertung der Daten aus einer Untersuchung von Hamouzova & Würthner (1976) dienen, die ein Kontrollgruppendesign zum Intelligenzcoaching des IST durchführten. Die Parameterschätzungen, die mit LISREL durchgeführt wurden, finden sich für das System (7.11) in Tabelle 7.1.

Während in der Gruppe A (intensives Training) jeweils der zum Posttest parallele Prätest der beste Einzelprädiktor ist, „zerfasert“ sich die Matrix  $A$  der Kontrollgruppe (s. Tabelle 7.1 rechts). Es scheint durch das Lernen von Denkstrategien eine Stabilisierung der Tests im korrelativen Sinne eingetreten zu sein (Tabelle 7.1, links). Zur Simulation der erwarteten Entwicklungsverläufe (Trajektorien) wird die Lösung des Differenzengleichungssystems (7.13) berechnet und geplottet (Fig. 7.1). Obwohl es bei zeitdiskreten Modellen unzulässig ist, haben wir der besseren Lesbarkeit halber die Entwicklungskurven durchgezogen. Es zeigen sich deutliche Unterschiede zwischen den beiden

Tabelle 7.1: Systemmatrizen  $A$  und Inputvektoren  $b$  für die Intensivtrainingsgruppe A und die Kontrollgruppe K (Die Parameter wurden im Rahmen einer multivariaten Regression der 9 I-S-T Posttests auf die 9 parallelen I-S-T Pretests geschätzt. Der jeweils größte Koeffizient einer Zeile wurde eingerahmt. Es erscheinen nur Koeffizienten vom Betrage größer als .10)

	Trainingsgruppe A									
Post- tests	Pretests									
	SE	WA	AN	GE	ME	RA	ZR	FA	WÜ	<i>b</i>
SE	.28	.13	.16	.18						5.08
WA	.15	.17	.12		.14		.11	.10		4.99
AN		.18	.95		-.26		.28	-.29		1.98
GE	.60	-.33	.42	.68	.56		-.36	-.15	-.31	2.83
ME	-.12		.22	-.14	.74			.39		1.34
RA		.32				.66			.14	-.69
ZR	.28	.26	.16		-.18		.53			-.29
FA	-.30	.14	.24					.55	.17	4.96
WÜ	.23	-.11						.17	.60	2.59

	Kontrollgruppe K									
Post- tests	Pretests									
	SE	WA	AN	GE	ME	RA	ZR	FA	WÜ	<i>b</i>
SE	.32			.16	.12	.12				5.06
WA	.13	.27	.39					.30	-.39	5.74
AN	.90	.11	.47	-.11	-.31	.19	.13	.15	-.26	-3.10
GE	.36		.70	.12	.17	-.11			-.19	4.87
ME			-.24	.34	.62	-.12	.12			.14
RA	.15	-.13	.23		.11	.77	.15			-.37
ZR	.26	-.10					.64	.16		.59
FA		-.16	.12				.17	.53	.27	2.31
WÜ	.17	-.36	.22	-.29			.34	.30	.48	2.69

Gruppen, die nach der Simulation mit Formel (7.13) frühestens nach 2 Monaten bzw. 24 Stunden verteiltem Training ( $8 \times 3$  Stunden) zu erwarten wären. Auf Grund der Simulation kann man vermuten, daß das Coaching effektiv (wenn auch verbesserungswürdig) aber zeitlich zu kurz angelegt war, um einen signifikanten Schereneffekt zwischen den Gruppen auf einigen Untertests deutlich werden zu lassen. Diese Vermutungen ließen sich auch an einer Längsschnittuntersuchung kreuzvalidieren (Möbus, 1981).

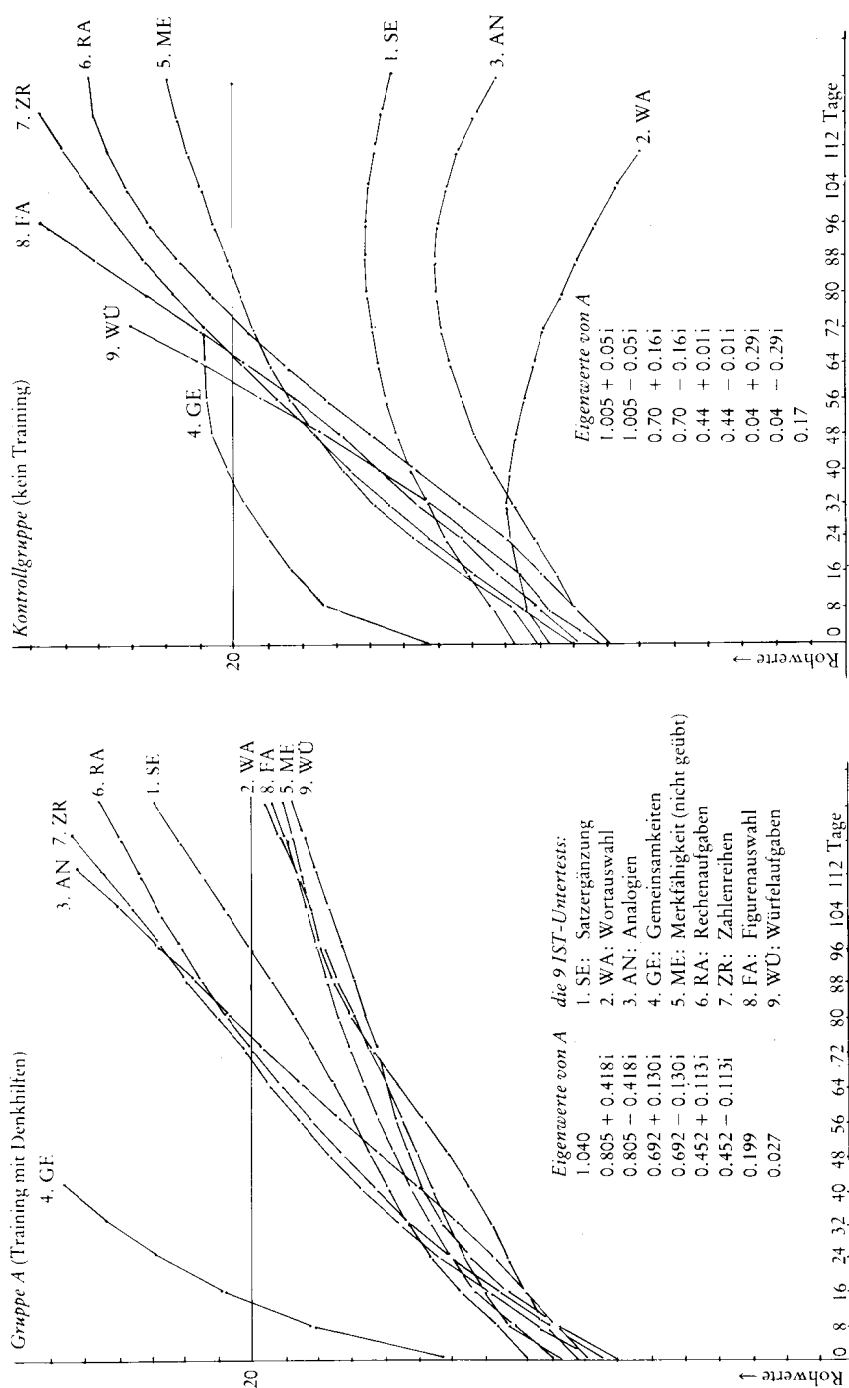


Fig. 7.1: Zwischen dem 1.-8. Tag beobachtete und dann anschließend mit Formel (7.13) simulierte Entwicklungsverläufe für die 9 I-S-T Untertests (SE = Satzergänzung, WA = Wortauswahl, AN = Analogien, GE = Gemeinsamkeiten, ME = Merkfähigkeit (wurde nicht trainiert), RA = Rechenaufgaben, ZR = Zahlenreihen, FA = Figurenauswahl, WÜ = Würfelaufgaben). Da bis auf GE alle Tests nur 20 Aufgaben besitzen, wurde bei  $X = 20$  eine waagerechte Linie eingezeichnet. Höhere Testleistungen wären nur zu erwarten, wenn man die Tests mit weiteren Aufgaben aufstocken würde.

## 7.4 Erwartungswertorientierte Analysen von Zeitreihen von Querschnitten: Schereneffekte bei Mittelwertsverläufen auf latenten Variablen

Bisher wurden nur die Mittelwertsverläufe auf manifesten Variablen untersucht. Dieses gilt auch für uni- wie auch multivariate Varianzanalysen. Ist man jedoch nicht an den Indikatormittelwerten sondern an Konstruktmittelwerten interessiert, führen die bisher behandelten Ansätze nicht zum Ziel. Diese Fragestellung läßt sich jedoch mit einem speziell erweitertem LISREL-Modell behandeln.

Dazu wird das ursprünglich von Jöreskog (1973) formulierte LISREL-Modell (7.1) und (7.2) nach einem Vorschlag von Sörbom (1979) erweitert zu

$$(7.16a) \quad \text{Strukturmodell} \quad B\eta = \alpha + \Gamma\xi + \zeta = [\alpha : \Gamma] \begin{bmatrix} 1 \\ \vdots \\ \xi \end{bmatrix} + [\zeta]$$

$$(7.16b) \quad 1. \text{ Meßmodell} \quad y = \nu_y + A_y\eta + \varepsilon = [\nu_y : A_y] \begin{bmatrix} 1 \\ \vdots \\ \eta \end{bmatrix} + [\varepsilon]$$

$$(7.16c) \quad 2. \text{ Meßmodell} \quad x = \nu_x + A_x\xi + \delta = [\nu_x : A_x] \begin{bmatrix} 1 \\ \vdots \\ \xi \end{bmatrix} + [\delta]$$

wobei  $\alpha$ ,  $\nu_y$  und  $\nu_x$  Vektoren mit Regressionskonstanten sind.

Es wird nicht mehr angenommen, daß  $E(\xi) = 0$  und  $E(\eta) = 0$  sind. Die Erwartungswerte der latenten Konstrukte der 1. Welle (Pretests) sind

$$(7.17a) \quad E(\xi) = \kappa$$

und die der 2. Welle (Konstrukte der Posttests)

$$(7.17b) \quad E(\eta) = B^{-1}[\alpha : \Gamma] \begin{bmatrix} 1 \\ \vdots \\ E(\xi) \end{bmatrix} + B^{-1}[E(\zeta)] = B^{-1}[\alpha : \Gamma] \begin{bmatrix} 1 \\ \vdots \\ \kappa \end{bmatrix}$$

und die der manifesten Variablen:

$$(7.17c) \quad E(x) = \nu_x + A_x \kappa = [\nu_x : A_x] \begin{bmatrix} 1 \\ \vdots \\ \kappa \end{bmatrix}$$

sowie

$$(7.17d)$$

$$E(y) = [\nu_y : A_y] \begin{bmatrix} 1 \\ \vdots \\ E(\eta) \end{bmatrix} + E(\varepsilon) = [\nu_y : A_y] \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ B^{-1}[\alpha : \Gamma] \begin{bmatrix} 1 \\ \vdots \\ \kappa \end{bmatrix} \end{bmatrix} = \nu_y + A_y B^{-1}(\alpha + \Gamma\kappa)$$

In einer *einigen* Population sind die Parameter  $\nu_y, \nu_x, \alpha, \kappa$  nicht identifiziert. Analysiert man jedoch simultan mehrere Gruppen, wie es z.B. bei einem Kontrollgruppendesign der Fall sein kann, gibt es einfache Restriktionen, die es ermöglichen, die Mittelwertparameter zu identifizieren (s. Sörbom, 1979). In der LISREL-Terminologie nimmt das Modell (7.16) die Struktur (7.18) oder noch allgemeiner (7.19) an. Dabei ändert sich die Parametermatrix  $\Sigma$  in (7.3) zur Momentenmatrix (um Null) der Indikatoren  $y$  und  $x$  sowie der Scheinvariablen 1:

(7.18a)

Strukturmodell

$$\begin{bmatrix} 1^* & 0^{**} & 0^{**} \\ 0^{**} & B & -\Gamma \\ 0^{**} & 0^{**} & I \end{bmatrix} \begin{bmatrix} 1 \\ \eta \\ \xi \end{bmatrix} = \begin{bmatrix} 1^* \\ \alpha \\ \kappa \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} + \begin{bmatrix} 0 \\ \zeta \\ \xi - E(\xi) \end{bmatrix}$$

(7.18b)

Meßmodell

$$\begin{bmatrix} z \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \nu_y & A_y & 0^{**} \\ \nu_x & 0^{**} & A_x \end{bmatrix} \begin{bmatrix} 1 \\ \eta \\ \xi \end{bmatrix} + \begin{bmatrix} \varepsilon \\ \delta \end{bmatrix}$$

oder noch allgemeiner:

(7.19a)

Strukturmodell

$$\begin{bmatrix} I^* & 0^{**} & -\nu_y & -A_y & 0^{**} \\ 0^{**} & I^* & -\nu_x & 0^{**} & -A_x \\ 0^{**} & 0^{**} & 1^* & 0^{**} & 0^{**} \\ 0^{**} & 0^{**} & -\alpha & B & -\Gamma \\ 0^{**} & 0^{**} & -\kappa & 0^{**} & I^* \end{bmatrix} \begin{bmatrix} y \\ x \\ 1 \\ \eta \\ \xi \end{bmatrix} = \begin{bmatrix} \varepsilon \\ \delta \\ 1 \\ \zeta \\ \xi - E(\xi) \end{bmatrix} \quad B\eta = \zeta$$

(7.19b)

Meßmodell

$$\begin{bmatrix} z \\ 1 \end{bmatrix} = \begin{bmatrix} y \\ x \\ 1 \end{bmatrix} = \begin{bmatrix} I^* & 0^{**} & 0^{**} & 0^{**} & 0^{**} \\ 0^{**} & I^* & 0^{**} & 0^{**} & 0^{**} \\ 0^{**} & 0^{**} & 1^* & 0^{**} & 0^{**} \end{bmatrix} \begin{bmatrix} y \\ x \\ 1 \\ \eta \\ \xi \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad y = A_y \eta$$

Die zu analysierende Matrix ist nicht mehr wie im ursprünglichen LISREL-Modell die Kovarianzmatrix  $S$  der Indikatoren sondern die Momentenmatrix der Indikatoren ( $z$  und der Scheinvariablen 1) um Null:

(7.20)

$$M = \begin{bmatrix} S + \bar{z}\bar{z}' & \bar{z} \\ \bar{z}' & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{N} \sum_{i=1}^N zz' & \frac{1}{N} \sum_{i=1}^N z \cdot 1 \\ \frac{1}{N} \sum_{i=1}^N 1 \cdot z' & \frac{1}{N} \sum_{i=1}^N 1 \cdot 1 \end{bmatrix}$$

wobei  $S$  = Kovarianzmatrix der  $z$  $S + \bar{z}\bar{z}'$  = Momentenmatrix der  $z$  um Null $\bar{z}$  = Mittelwertsvektor der  $z$  = Produkt-Moment der  $z$   
mit der Scheinvariablen 1

Wir wollen jetzt die Fragestellung: „Wie verändern sich Gruppenunterschiede zwischen Konstruktmittelwerten zu diskreten Zeitpunkten?“ mit LISREL behandeln. Dazu nehmen wir das Beispiel von Jöreskog und Sörbom (1976). Es soll untersucht werden, ob sich zwischen den beiden Gruppen ein Schereneffekt herausbildet. Jeweils 2 Untertests des Sequential Test of Educational Progress (STEP) dienen als Indikatoren für die latenten Variablen  $\xi$  (1. Welle) und  $\eta$  (2. Welle). Die Kovarianzmatrizen und Mittelwerte finden sich in Tabelle 7.2. Nach (7.20) lassen sich die Momentmatrizen um Null berechnen, die dann dem LISREL-Modell (7.18) für die Maximum-Likelihoodschätzungen als Daten dienen (Tabelle 7.3). Für jede der beiden Gruppen wird ein Modell (7.18) aufgestellt, dabei werden folgende Restriktionen eingeführt: (1) die Parameter des Meßmodells sind für beide Gruppen gleich, (2) die Parameter  $\alpha$  der Bezugsgruppe („Jungen mit akademischem Curriculum“) werden Null gesetzt: damit sind die Mittelwerte der latenten Variablen  $\xi_1$  und  $\eta_1$  der Bezugsgruppe in den Nullpunkt verschoben. Die Einschränkung (1) bedeutet nicht die Gleichsetzung der Kovarianzmatrizen sondern ermöglicht es erst, indikatorspezifische Mittelwertsunterschiede im Konstruktraum abzubilden. Die Parameterschätzungen finden sich in Tabelle 7.4.

Es lassen sich zwei gruppenspezifische Regressionen im Raum der latenten Variablen schätzen (s.a. Figur 7.2)

$$(7.21a) \quad \text{für Jungen mit „akademischem“ Curriculum:} \\ \eta_1 = .963 \xi_1 + \zeta_1$$

$$(7.21b) \quad \text{für Jungen mit „nichtakademischem“ Curriculum:} \\ \eta_2 = -6.121 + .811 \xi_1 + \zeta_2$$

Die Mittelwerte auf den latenten Variablen sind

$$(7.22a) \quad \text{für Jungen mit „akademischem“ Curriculum:} \\ \xi_1 = 0.0^* \quad \eta_1 = 0.0^*$$

(7.22b) für Jungen mit „nichtakademischem“ Curriculum:  
 $\xi_2 = -13.75 \quad \bar{\eta}_2 = -17.27$

Es zeigt sich, daß die Mittelwertsdifferenz *auf den latenten Variablen* größer wird. Signifikanztests lassen sich durchführen, wenn man durch entsprechende Restriktionen, das volle Modell reduziert und die beiden Likelihoods mit einem Likelihoodquotiententest vergleicht. Will man z.B. die Hypothese testen, daß kein Schereneffekt vorliegt, kann man die Restriktion testen, daß die Steigungen der beiden Regressionen gleich sein sollen.

Das Modell ist im Grunde ein statisches Modell, weil man im Gegensatz zum vorher behandelten Differenzengleichungssystem, keine expliziten Extrapolations- und Prognosevorschriften zur Hand hat. Hier wird ganz deutlich, daß das Modell aus der regressions- und faktorenanalytischen Denkschule der Statistik stammt. Andererseits ist es möglich, die dynamischen zeitdiskreten oder zeitkontinuierlichen Systeme auf latente Variablen auszudehnen. Hierzu müssen die Outputmatrizen  $H$  in (9.1b) und (9.2b) und die Inputmatrizen  $B$  bestimmte Strukturen annehmen. So darf  $H$  nicht gleich  $I$  sein, sondern muß ähnlich wie die Lambda-Matrizen in den LISREL-Meßmodellen spezifiziert sein. Sinnvolle Modelle müssen auch hier bestimmten Kriterien genügen. So

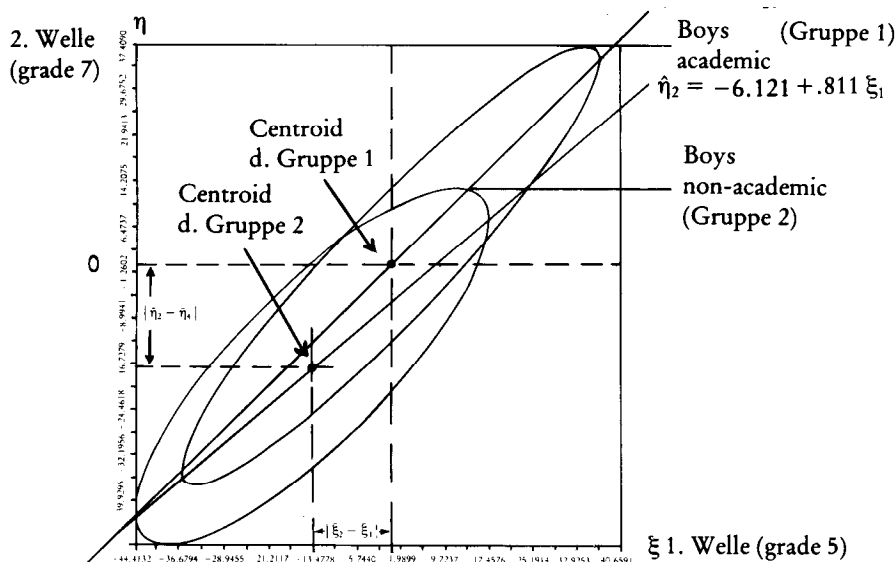


Fig. 7.2: Raum der latenten Variablen für die 1. und 2. Welle: Regressionen der latenten Variablen der 2. Welle auf die 1. Welle für die beiden Curriculumgruppen



sind Fragen der Beobachtbarkeit und Kontrollierbarkeit zu untersuchen (s.a. Athans et al., 1974).

Tabelle 7.2: Stichprobenkovarianzmatrizen für zwei Curricula

Boys academic (N = 373)				
STEP reading, grade 5 $X_1$	281.349			
STEP writing, grade 5 $X_2$	184.219	182.821		
STEP reading, grade 7 $Y_1$	216.739	171.699	283.289	
STEP writing, grade 7 $Y_2$	198.376	153.201	208.837	246.069
Boys non-academic (N = 249)				
STEP reading, grade 5 $X_1$	174.485			
STEP writing, grade 5 $X_2$	134.468	161.869		
STEP reading, grade 7 $Y_1$	129.840	118.836	228.449	
STEP writing, grade 7 $Y_2$	102.194	97.767	136.058	180.460
Mittelwerte				
	Boys academic		Boys non-academic	
STEP reading, grade 5 $X_1$	262.236		248.675	
STEP writing, grade 5 $X_2$	258.788		246.896	
STEP reading, grade 7 $Y_1$	275.630		258.546	
STEP writing, grade 7 $Y_2$	269.075		253.349	

Tabelle 7.3 : Zu analysierende Momentenmatrizen M

	Gruppe 1: „akademisches“ Curriculum					Gruppe 2: „nichtakad.“ Curriculum				
	$X_1$	$X_2$	$Y_1$	$Y_2$	Konst.	$X_1$	$X_2$	$Y_1$	$Y_2$	Konst.
$X_1$	69029	.	.	.	.	62025	.	.	.	.
$X_2$	68041	67159	.	.	.	61537	61120	.	.	.
$Y_1$	72478	71496	76237	.	.	64443	63966	67101	.	.
$Y_2$	70755	69795	74372	72660	.	63097	62637	65638	64340	.
1	262	258	275	269	1	248	246	258	253	1

Tabelle 7.4: Parameterschätzungen mit LISREL

Gruppe 1: „akademisches“ Curriculum

Strukturmodell

$$\begin{bmatrix} 1^* & 0^* & 0^* \\ 0^* & 1^* & 0^* \\ 0^* & -\Gamma & B \end{bmatrix} \begin{bmatrix} 1 \\ \xi \\ \eta \end{bmatrix} = \begin{bmatrix} 1^* \\ \kappa \\ \alpha \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} + \begin{bmatrix} 0 \\ \xi - E(\xi) \\ \zeta \end{bmatrix}$$

mit  $I = 1^*$ ,  $-\Gamma = -.963$ ,  $B = 1^*$ ,  $\kappa = 0^*$ ,  $\alpha = 0^*$

Meßmodell

$$\begin{bmatrix} z \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} v_{x_1} = 262 & 1^* & 0^* \\ v_{x_2} = 258 & \lambda_{x_2} = .839 & 0^* \\ v_{y_1} = 275 & 0^* & 1^* \\ v_{y_2} = 269 & 0^* & \lambda_{y_2} = .895 \end{bmatrix} \begin{bmatrix} 1 \\ \xi \\ \eta \end{bmatrix} + \begin{bmatrix} \delta \\ \varepsilon \end{bmatrix}$$

$$z = \begin{bmatrix} v_x & \lambda_x & 0^* \\ v_y & 0^* & \lambda_y \end{bmatrix} \begin{bmatrix} 1 \\ \xi \\ \eta \end{bmatrix} + \begin{bmatrix} \delta \\ \varepsilon \end{bmatrix}$$

Gruppe 2: „nichtakademisches“ Curriculum

Strukturmodell

$$\begin{bmatrix} 1^* & 0^* & 0^* \\ 0^* & 1^* & 0^* \\ 0^* & -.811 & 1^* \end{bmatrix} \begin{bmatrix} 1 \\ \xi \\ \eta \end{bmatrix} = \begin{bmatrix} 1^* \\ -13,75 \\ -6,121 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} + \begin{bmatrix} 0 \\ \xi - E(\xi) \\ \zeta \end{bmatrix}$$

## 8. Markoff-Modelle für qualitative Variable bei diskreter Zeit

Während bei den quantitativen Variablen die Variablenwerte zu verschiedenen Zeitpunkten (oder daraus abgeleitete andere Maße wie Differenzen oder Verhältnisse) als Operationalisierung von Veränderungen gelten, betrachtet man bei qualitativen Variablen analog dazu Veränderungen in den Wahrscheinlichkeiten, sich in bestimmten Zuständen zu befinden (s. Coleman, 1968). Die dabei verwendeten Modelle sind meistens die zeitdiskreten Markoff-Ketten, die in die Klasse der Differenzgleichungssysteme einzuordnen sind. Anwendungen finden sich bei der Beschreibung von Konditionierungs-, Konzeptlern-, Informationsspeicherungs-, Denk- und Interaktionsprozessen (Atkinson & Estes, 1963; Estes & Suppes, 1974; Deppe, 1977; Fararo, 1973; Greeno, 1974; Harder, 1973; Laming, 1973; Levine & Burke, 1972; Massy et al., 1970; Miller & Chomsky, 1963; Rapoport, 1963; Restle & Greeno, 1970; Revenstorf et al., 1977; Tack, 1976).

Bevor wir Markoff-Ketten 1. und 2. Ordnung behandeln, wollen wir einige Grundbegriffe einführen: (a) bedingte Wahrscheinlichkeiten, (b) Markoffannahme, (c) Stationarität.

### (a) Bedingte Wahrscheinlichkeit

Die Variable  $Y$  habe  $K$  Ausprägungen (bzw. Zustände). Es werden mehrere Zeitpunkte betrachtet. Die Verteilung der Variablen zum Zeitpunkt  $t$  sei abgekürzt durch einen Vektor dargestellt:

$$(8.1) \quad p'(t) = (p_1(t), \dots, p_K(t))$$

wobei  $p_i(t)$  die Wahrscheinlichkeit ist, daß die Variable  $Y$  zum Zeitpunkt  $t$  den Wert  $i$  annimmt ( $p_i(t) = W(\{Y_t = i\})$ )

Die bedingten Wahrscheinlichkeiten können folgendermaßen abgekürzt geschrieben werden:

$$(8.2) \quad p_{ij}(t_1, t_2) = W(\{Y_{t_2} = j\} \mid \{Y_{t_1} = i\}) = \text{Wahrscheinlichkeit, daß die Variable } Y \text{ zum Zeitpunkt } t_2 \text{ den Wert } j \text{ annimmt unter der Bedingung, daß } Y \text{ zum Zeitpunkt } t_1 \text{ den Wert } i \text{ besaß.}$$

Die verschiedenen bedingten Wahrscheinlichkeiten kann man übersichtlich in einer Matrix zusammenfassen:

$$(8.3) \quad P(t_1, t_2) = \begin{bmatrix} p_{11}(t_1, t_2) & \dots & p_{1K}(t_1, t_2) \\ \dots & \dots & \dots \\ p_{K1}(t_1, t_2) & \dots & p_{KK}(t_1, t_2) \end{bmatrix}$$

Die Summe pro Zeile ist jeweils 1, denn es ist

$$\sum_{j=1}^K W(\{Y_{t_2} = j\} \mid \{Y_{t_1} = i\}) = 1$$

Nach dem Satz für die totale Wahrscheinlichkeit (s. z.B. Storm, 1969) kann man aufgrund der Kenntnis der Wahrscheinlichkeiten zum Zeitpunkt  $t_1$  und der bedingten Wahrscheinlichkeiten  $p_{ij}(t_1, t_2)$  die Wahrscheinlichkeiten für den Zeitpunkt  $t_2$  berechnen:

$$(8.4) \quad p_i(t_2) = W(\{Y_{t_2} = j\}) = \sum_{i=1}^K W(\{Y_{t_2} = j\} \mid \{Y_{t_1} = i\}) W(\{Y_{t_1} = i\}) = \sum_{i=1}^K p_{ij}(t_1, t_2) p_i(t_1)$$

Faßt man (8.4) für jedes  $j$  zusammen, ergibt sich folgende Matrixmultiplikation:

(8.5)

$$\boxed{p'(t_2)} = \boxed{p'(t_1)} \cdot \boxed{P(t_1, t_2)}$$

*In den bedingten Wahrscheinlichkeiten sind alle Informationen enthalten, die die Veränderung zwischen zwei Zeitpunkten charakterisieren.*

Manchmal werden auch zwei Prozesse miteinander kombiniert, so daß nur noch eine einzige Übergangsmatrix  $P(t_1, t_2)$  vorliegt (Greeno, 1974). So kann die Variable Y beim einfachen Alles-oder-nichts-Lernen zwei Ausprägungen besitzen: S = Information gespeichert, U = Information nicht im Gedächtnis gespeichert. Die Übergangsmatrix des Lernprozesses könnte folgendermaßen aussehen:

$$P_1(t_1, t_2) = \begin{array}{l} \text{Information ge-} \\ \text{speichert } S(t_1) \\ \text{Information nicht} \\ \text{im Gedächtnis ge-} \\ \text{speichert } U(t_1) \end{array} \begin{array}{cc} S(t_2) & U(t_2) \\ \hline 1 & 0 \\ \hline a & (1 - a) \end{array}$$

S und U sind dabei latente, nicht direkt beobachtbare Zustände. a repräsentiert einen unbekannten Parameter, der durch Daten geschätzt werden muß.

Aus den Antworten einer Person ist aber nicht ein direkter Schluß auf die latenten Zustände S, U der Person möglich. Sie kann trotz nicht gespeicherter Information richtige Reaktionen (z. B. Raten in einem multiple choice Fragebogen) zeigen. Dieser zweite Prozeß (Performanzmodell) besitzt ebenfalls eine Übergangsmatrix

$$P_2(t_1, t_2) = \begin{array}{l} \text{Information ge-} \\ \text{speichert } S(t_1) \\ \text{Information nicht} \\ \text{gespeichert } U(t_1) \end{array} \begin{array}{cc} \text{richtige} & \text{falsche} \\ \text{Reaktion } C(t_2) & \text{Reaktion } E(t_2) \\ \hline 1 & 0 \\ \hline p & (1 - p) \end{array}$$

C und E sind jetzt direkt beobachtbare manifeste Zustände. Beide Prozesse lassen sich kombinieren mit neuen Zuständen SC, UC und UE. Dabei kann

nur das Auftreten von UE direkt über die nicht korrekten Antworten einer VP beobachtet werden. Die Übergangsmatrix des kombinierten Lern-Performanzmodells ist

	SC( $t_2$ )	UC( $t_2$ )	UE( $t_2$ )
	Speicherung und korrekte Reaktion	keine Speicherung und korrekte Reaktion	keine Speicherung und Fehler (nicht korrekte Reaktion)
$P(t_1, t_2) =$	SC( $t_1$ )	UC( $t_1$ )	UE( $t_1$ )
	1	0	0
	a	$(1 - a)p$	$(1 - a)(1 - p)$
	a	$(1 - a)p$	$(1 - a)(1 - p)$

Die Modelle mit latenten Zuständen werfen eine Reihe von Problemen auf, die unter 8.8.2 noch gesondert besprochen werden.

Auf ähnliche Weise wie in (8.5) wird die Zustandsverteilung des 2. Zeitpunkts in die des 3. Zeitpunkts transformiert:

$$(8.6) \quad p'(t_3) = p'(t_2) \cdot P(t_2, t_3)$$

Durch Substitution von (8.5) für  $p'(t_2)$  in (8.6) erhält man dann

$$(8.7) \quad p'(t_3) = p'(t_1) \cdot P(t_1, t_2)P(t_2, t_3)$$

Allgemein bekommt man durch wiederholtes Einsetzen für mehrere Zeitpunkte ( $t_1 < t_2 < t_3 \dots < t_m$ )

$$(8.8) \quad p'(t_m) = p'(t_1) \cdot P(t_1, t_2) \cdot P(t_2, t_3) \cdots P(t_{m-1}, t_m) = p'(t_1) \cdot \prod_{i=2}^m P(t_{i-1}, t_i)$$

Andererseits könnte man auch die Wahrscheinlichkeiten zum Zeitpunkt  $t_m$  aus den Wahrscheinlichkeiten  $p'(t_1)$  und den bedingten Wahrscheinlichkeiten  $P(t_1, t_m)$  erhalten:

$$(8.9) \quad p'(t_m) = p'(t_1) \cdot P(t_1, t_m)$$

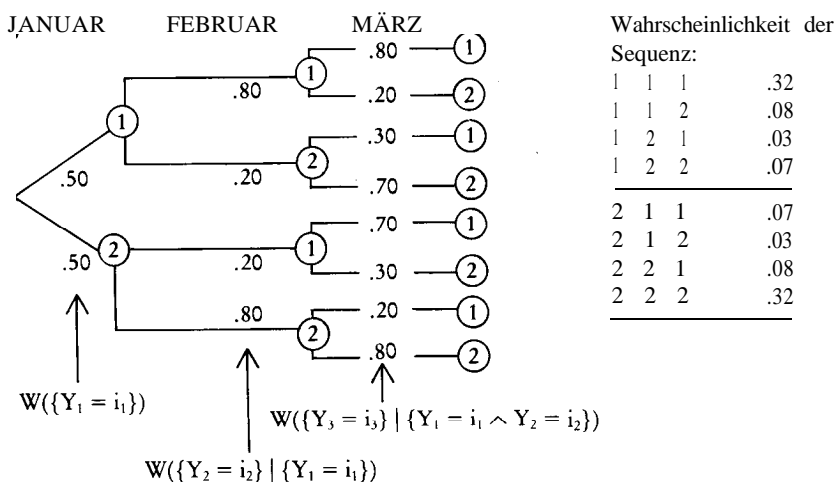
Auf den ersten Blick würde man vermuten, daß  $P(t_1, t_m)$  gleich dem Produkt  $\prod_{i=2}^m P(t_{i-1}, t_i)$  sein müßte. Dieses gilt nicht allgemein. Das soll an folgendem

Beispiel gezeigt werden.

*Beispiel:*

Die untersuchte Variable Y habe zwei Ausprägungen: 1 = gesund geschrieben, 2 = krank geschrieben. Aufgrund jahrelanger Erfahrung wisse man, daß nach

einer im Dezember durchgeführten Therapie für die drei Zeitpunkte Januar ( $t_1$ ), Februar ( $t_2$ ) und März ( $t_3$ ) folgende Wahrscheinlichkeiten gelten, gesund bzw. krank geschrieben worden zu sein:



Aus diesen Wahrscheinlichkeiten kann man dann die Wahrscheinlichkeiten für die einzelnen Sequenzen berechnen:

$$\begin{aligned}
 W(\{Y_1 = i_1 \wedge Y_2 = i_2 \wedge Y_3 = i_3\}) &= W(\{Y_3 = i_3\} \mid \{Y_1 = i_1 \wedge Y_2 = i_2\}) \\
 &\quad \cdot W(\{Y_2 = i_2\} \mid \{Y_1 = i_1\}) \\
 &\quad \cdot W(\{Y_1 = i_1\})
 \end{aligned}$$

Die Wahrscheinlichkeit für den 1. Ast des Wahrscheinlichkeitsbaumes (Sequenz: 1 | 1 | 1) ist:

$$W(\{Y_1 = 1 \wedge Y_2 = 1 \wedge Y_3 = 1\}) = .80 \cdot .80 \cdot .50 = .32$$

Die Matrizen für die bedingten Wahrscheinlichkeiten kann man aus den Werten für die Sequenzen berechnen. Man erhält als Matrix der bedingten Wahrscheinlichkeiten für den Übergang von Januar auf Februar:

$$P(t_1, t_2) = \begin{matrix} & \begin{matrix} \textcircled{1} & \textcircled{2} \end{matrix} \\ \begin{matrix} \textcircled{1} \\ \textcircled{2} \end{matrix} & \begin{bmatrix} .80 & .20 \\ .20 & .80 \end{bmatrix} \end{matrix}$$

und für den Übergang von Februar auf März:

$$P(t_2, t_3) = \begin{matrix} & \textcircled{1} & \textcircled{2} \\ \textcircled{1} & \begin{bmatrix} .78 & .22 \end{bmatrix} \\ \textcircled{2} & \begin{bmatrix} .22 & .78 \end{bmatrix} \end{matrix} = \begin{bmatrix} .80 \cdot .80 + .20 \cdot .70 & .20 \cdot .30 + .80 \cdot .20 \\ .80 \cdot .20 + .20 \cdot .30 & .20 \cdot .70 + .80 \cdot .80 \end{bmatrix}$$

Nach Formel (8.7) ist dann damit die Verteilung  $p'(t_3)$  (Wahrscheinlichkeiten im März, gesund bzw. krank geschrieben zu sein):

$$\begin{aligned} p'(t_1) \cdot P(t_1, t_2) \cdot P(t_2, t_3) &= \\ = (.5, .5) \cdot \begin{bmatrix} .80 & .20 \\ .20 & .80 \end{bmatrix} \cdot \begin{bmatrix} .78 & .22 \\ .22 & .78 \end{bmatrix} &= (.5, .5) \cdot \boxed{\begin{bmatrix} .668 & .332 \\ .332 & .668 \end{bmatrix}} = (.5, .5) = p'(t_3) \end{aligned}$$

oder direkt nach Formel (8.9):

$$p'(t_1) \cdot P(t_1, t_3) = (.5, .5) \cdot \boxed{\begin{bmatrix} .70 & .30 \\ .30 & .70 \end{bmatrix}} = (.5, .5) = p'(t_3)$$

Die beiden Matrizen (im Beispiel doppelt angestrichen) stimmen nicht überein. Das Produkt von Matrizen mit bedingten Wahrscheinlichkeiten ist nicht selbst die entsprechende Matrix der bedingten Wahrscheinlichkeiten. Das trifft nur dann zu, wenn die sogenannte „Markoffannahme“ (siehe unten) erfüllt ist.

### b) Markoffannahme und einige Bezeichnungen

Bei bedingten Wahrscheinlichkeiten gilt im allgemeinen (z.B. bei 3 Zeitpunkten) (8.10) nicht. Falls die Gleichheit gefordert wird, spricht man bei diesem Beispiel von „Markoffannahme 1. Ordnung“. Diese Annahme vereinfacht die theoretische Analyse der Prozesse und deren Schätzbarkeit erheblich.

$$(8.10) \quad W(\{Y_t = i_t\} \mid \{Y_{t-2} = i_{t-2}, Y_{t-1} = i_{t-1}\}) = W(\{Y_t = i_t\} \mid \{Y_{t-1} = i_{t-1}\})$$

im obigen Beispiel ist für  $i_t = 1, i_{t-1} = 1, i_{t-2} = 1$ :

$$\begin{aligned} W(\{Y_3 = 1\} \mid \{Y_1 = 1, Y_2 = 1\}) &= .80 \neq \\ W(\{Y_3 = 1\} \mid \{Y_2 = 1\}) &= .78 \end{aligned}$$

Man müßte bei mehreren Zeitpunkten die ganze Vergangenheit des Prozesses mitberücksichtigen; dadurch ergäbe sich bei lang dauernden Prozessen eine immense Fülle von bedingten Wahrscheinlichkeiten. Darüber hinaus wären auch Prognosen unmöglich, da für jede Vergangenheit (spezielle Sequenz von Ausprägungen) völlig neue unbekannte bedingte Wahrscheinlichkeiten für die Prognosen benötigt würden.

Die Markoffeigenschaft eines Prozesses erlaubt es, nicht die gesamte Vergangenheit berücksichtigen zu müssen. Falls dann die Gleichung (8.10) gilt, wäre damit die Markoweigenschaft 1. Ordnung gegeben, da die Wahrscheinlichkeit, welchen Zustand der Prozeß zum Zeitpunkt  $t$  annimmt, nur vom unmittelbar vorher stattfindenden Ereignis abhängt.

Ein Prozeß *k-ter Ordnung* wird durch folgende Forderung charakterisiert:

$$(8.11) \quad \begin{aligned} W(\{Y_t = i_t\} \mid \{Y_1 = i_1, Y_2 = i_2, \dots, Y_{t-k}, \dots, Y_{t-1} = i_{t-1}\}) = \\ W(\{Y_t = i_t\} \mid \{Y_{t-k} = i_{t-k}, \dots, Y_{t-1} = i_{t-1}\}) \end{aligned}$$

Dabei sei  $Y_1$  der Wert zu Beginn des Prozesses. Die Wahrscheinlichkeit, welchen Zustand  $i_t$  der Prozeß zum  $t$ -ten Zeitpunkt annimmt, hängt nur vom Verlauf des Prozesses während der letzten  $k$  Zeitpunkte ab. Frühere Zustände spielen keine Rolle mehr. Prozesse, die für diskrete Zeitpunkte betrachtet werden, heißen mit dieser Annahme *Markoffketten k-ter Ordnung*.

c) Falls zusätzlich die bedingten Wahrscheinlichkeiten für alle Zeitpunkte  $t$  gleich sind, spricht man von einer homogenen (auch stationären) Markoffkette.

Ein Spezialfall ist der unabhängige Prozeß (Markoffkette 0-ter Ordnung), bei dem gilt (independent trial process):

$$W(\{Y_t = i_t\} \mid \{Y_1 = i_1, Y_2 = i_2, \dots, Y_{t-1} = i_{t-1}\}) = W(\{Y_t = i_t\})$$

Dabei hängt das Ergebnis zum Zeitpunkt  $t$  überhaupt nicht von der Vergangenheit ab.

## 8.1 Markoffketten 1. Ordnung mit einer Variablen

Man kann wieder wie bei Gleichung (8.8) die Verteilung für einen bestimmten Zeitpunkt  $t$  mit Hilfe der Übergangswahrscheinlichkeiten und der Anfangsverteilung berechnen.

$$(8.12) \quad p'(t) = p'(1) \prod_{s=2}^t P(s-1, s)$$



Andererseits gilt aber auch:

$$(8.13) \quad p'(t) = p'(1) P(1, t)$$

Infolge der Markoweigenschaft stimmt aber nun  $P(1, t)$  mit dem Produkt:

$\prod_{s=2} P(s-1, s)$  überein\*. Zudem kann man jetzt für *theoretische* Analysen die

Anfangsverteilung frei wählen und so Konsequenzen unterschiedlicher Anfangsverteilungen simulieren.

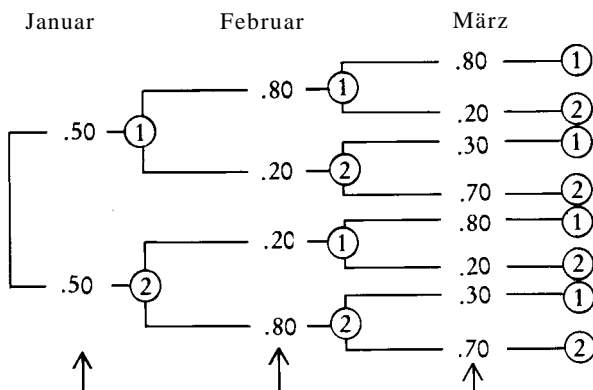
Falls der Prozeß (Zeit-)homogen ist, wird die Gleichung (8.12) noch einfacher. Denn dann gilt:  $P(s-1, s) = P(t-1, t) = P$  und damit

$$(8.14) \quad p'(t) = p'(1) \underbrace{P \dots P}_{2 \dots t} = p'(1) P^{(t-1)}$$

Diese Formel weist Ähnlichkeit zur Berechnung der „Lösung“ (= Trajektorie, Erwartungswertverläufe) von Differenzengleichungssystemen ohne exogenen Input auf (siehe 7.13)

Fortsetzung des Beispiels mit 2 Modifikationen:

Falls der „Wahrscheinlichkeitsbaum“ monatlicher jeweils eintägiger therapeutischer Behandlung wie in der nachfolgenden Figur aussähe, läge Markoff-eigenschaft 1. Ordnung vor.



$$W(\{Y_1 = i_1\}), W(\{Y_2 = i_2\} \mid \{Y_1 = i_1\}) \quad W(\{Y_3 = i_3\} \mid \{Y_1 = i_1, Y_2 = i_2\})$$

\* Damit kann man Übergangswahrscheinlichkeiten zwischen weiter auseinanderliegenden Zeitpunkten durch Multiplikation der Matrizen der Übergänge von benachbarten Zeitpunkten gewinnen.

Die Zustandswahrscheinlichkeiten für den März sind je nach Gesundheitszustand im Februar verschieden. Jedoch wirkt sich der jeweilige Gesundheitszustand im Januar nicht mehr auf die Wahrscheinlichkeiten im März aus. Man kann diesen Sachverhalt auch noch anders formulieren. Man kann z.B. folgende Wahrscheinlichkeit berechnen:

$$W(\{Y_3 = i_3\} \mid \{Y_2 = i_2\})$$

$$\text{z.B.: } W(\{Y_3 = 1\} \mid \{Y_2 = 1\}) = .80$$

Da diese (und auch die anderen Möglichkeiten) jeweils gleich

$W(\{Y_3 = 1\} \mid \{Y_1 = 1, Y_2 = 1\}) = .80$   
 und  $W(\{Y_3 = 1\} \mid \{Y_1 = 2, Y_2 = 1\}) = .80$   
 sind, ist die Markoffeigenschaft erster Ordnung gegeben (bei nur 3 Zeitpunkten lassen sich höhere Ordnungen nicht mehr überprüfen)

Als Übergangsmatrizen hat man:

$$P(2, 3) = {}_{t_2} \begin{array}{c|c} & t_3 \\ \hline & \begin{array}{cc} .80 & .20 \\ .30 & .70 \end{array} \end{array}$$

$$P(1, 2) = {}_{t_1} \begin{array}{c|c} & t_2 \\ \hline & \begin{array}{cc} .80 & .20 \\ .20 & .80 \end{array} \end{array}$$

Der Prozeß ist nicht homogen, da  $P(2,3) \neq P(1,2)$  mit Anfangsverteilung  $p'(1) = (.5, .5)$  ist.

Der Prozeß soll nun noch weiter durch die Annahme modifiziert werden, daß auch  $P(1,2)$  gleich  $P(2,3)$  ist; zudem sei angenommen, daß der Prozeß mit der gleichen Übergangsmatrix in derselben Weise fort dauert:

$$P_{(t, t-1)} = \begin{array}{c|c} \textcircled{1} & \textcircled{2} \\ \hline \textcircled{1} & \begin{array}{cc} .80 & .20 \\ .30 & .70 \end{array} \\ \textcircled{2} & \end{array}$$

Würde folgende Anfangsverteilung gelten:

$$p(0) = (1, 0)$$

würden alle Personen im Januar gesund geschrieben sein.

Ist für die darauffolgenden Monate entsprechend der angegebenen Übergangswahrscheinlichkeiten mit Krankheitsfällen zu rechnen, kann man den Anteil der verbliebenen „Gesunden“ für jeden Monat herleiten:

Abnahme der Wahr-  
scheinlichkeit,  
„gesund“ „krank“ „gesund“ zu sein

(8.15)

1    0	.80   .20 .30   .70	=	.80   .20	.20	Februar
.80   .20	.80   .20 .30   .70	=	.70   .30	.10	März
.70   .30	.80   .20 .30   .70	=	.65   .35	.05	April
.65   .35	.80   .20 .30   .70	=	.625   .375	.025	Mai

Diese Ergebnisse hätten aber auch nach (8.14) über die Potenzen von  $P$  berechnet werden können:

$$P^2 = \begin{bmatrix} .70 & .30 \\ .45 & .55 \end{bmatrix}$$

$$P^3 = \begin{bmatrix} .65 & .35 \\ .525 & .475 \end{bmatrix}$$

$$P^4 = \begin{bmatrix} .625 & .375 \\ .5625 & .4375 \end{bmatrix}$$

Die Verteilung für Mai ist:  $p'(4) = p'(0) \cdot P^4 = \begin{bmatrix} .625 & .375 \end{bmatrix}$

Bei näherer Betrachtung der Potenzen von  $P$  fällt auf, daß pro Spalte die Differenzen zwischen den Übergangswahrscheinlichkeiten bei zunehmenden Exponenten immer kleiner werden. Daher liegt es nahe, zu fragen, ob bei zunehmendem Potenzieren eine Übergangsmatrix mit gleichen Zeilen resultiert.

Im Beispiel ist:

$$P^5 = \begin{bmatrix} .6062 & .3938 \\ .5906 & .4094 \end{bmatrix}$$

$$\text{oder } P^8 = \begin{bmatrix} .5977 & .4023 \\ .6016 & .3984 \end{bmatrix}$$

Schon bei einer Potenz von 8 erhält man eine Matrix, deren Zeilen fast gleich  $\begin{bmatrix} .60 & .40 \end{bmatrix}$  sind.

Allgemein gilt für Übergangsmatrizen, die bei irgendeiner Potenz nur noch positive Elemente haben, (so daß der Übergang von jedem zu jedem anderen Zustand möglich ist):

$$(8.16) \quad \lim_{t \rightarrow \infty} P^t = \begin{bmatrix} \pi' \\ \vdots \\ \pi' \end{bmatrix} \quad \text{wobei } \pi' \text{ ein Zeilenvektor ist} \\ \text{mit } \sum_{i=1}^K \pi_i = 1 \text{ und } 0 < \pi_i < 1$$

Die Kenntnis des gegenwärtigen Zustands des Prozesses bietet keine zusätzliche Information für die Prognose eines in weiter Zukunft liegenden Zustandes (Für jemand, der jetzt gesund geschrieben ist, gilt die gleiche Wahrscheinlichkeit, in ferner Zukunft gesund geschrieben zu sein wie für einen jetzt gerade krank geschriebenen).

Ferner stellt der Vektor  $\pi'$  die Verteilung dar, die ein konstant wirkender Prozeß nicht mehr ändert (stationäre Verteilung, bzw. Gleichgewichtszustand).

Man kann zeigen (Ferschl, 1970), daß

$$(8.17) \quad \pi' = \pi' P$$

ist.

Für das Beispiel ist:  $\pi' = (.60 \ .40)$

$$\begin{bmatrix} .60 & .40 \end{bmatrix} \begin{bmatrix} .80 & .20 \\ .30 & .70 \end{bmatrix} = \begin{bmatrix} .60 & .40 \end{bmatrix}$$

Aus der Relation (8.17) läßt sich die stationäre Verteilung auch rechnerisch (unter Berücksichtigung, daß  $\sum \pi_i = 1$  ist) sehr einfach bestimmen. Es gibt noch eine dritte Interpretation für  $\pi$ . Läuft ein solcher Prozeß ab, verweilt eine Person einen bestimmten Anteil der Zeitpunkte in einem der Zustände.  $\pi$  gibt dann (bei  $t \rightarrow \infty$ ) den Anteil der Zeitpunkte an, zu denen sich die Person im Zustand  $i$  aufgehalten hat.

Eine weitere Interpretation des Prozesses mit Hilfe von  $\pi$  bezieht sich auf die „Erstrückkehrzeit“. Darunter versteht man die Zeit  $s$ , die zwischen dem Verlassen des Zustandes  $i$  zum Zeitpunkt  $t$  und der Rückkehr nach  $i$  zum Zeitpunkt  $t+s$  verstreicht.

Die mittlere Rückkehrzeit für den Zustand  $i$  ist:

$$\mu_i = \frac{1}{\pi_i}$$

Für das Beispiel ist:

$$\mu_G = \frac{1}{.60} = 1.66$$

$$\mu_K = \frac{1}{.40} = 2.50$$

d.h.: falls jemand krank geschrieben wird, dauert es im Schnitt 1.66 Monate bis er wieder gesund wird.

Wir haben mit diesen Interpretationen von  $\pi$  einige wichtige Aspekte herausgegriffen. Cox, D. R. & Miller, H. D. (1968), Kemeny & Snell (1965), Ferschl (1970) geben weitere Details und Beweise.

In der Literatur sind verschiedene *Arten von Markoffketten (bzw. Zuständen)* bekannt geworden.

#### a) Reguläre oder ergodische Markoffketten

Kemeny & Snell (1965) bezeichnen solche Markoffketten als regulär, bei denen Übergänge von allen zu allen Zuständen bei irgendeiner Potenz von  $P$  möglich sind.

Die Aussage (8.16) gilt nur für reguläre Ketten. In unserem Beispiel ist schon für die erste Potenz der Matrix  $P$  (also für  $P$  selbst) jeder Übergang möglich. Deshalb ist diese Markoffkette regulär.

#### b) Absorbierende Ketten

Falls mindestens ein Zustand existiert, der zwar erreicht, aber nicht mehr verlassen werden kann, heißt die Kette absorbierend und der Zustand selbst ebenso.

Gerade bei bestimmten theoriegeleiteten Modellen treten solche absorbierende Zustände immer wieder auf. Etwa in der Lerntheorie (siehe Bower & Trabasso (1964), Simon & Newell (1974) oder Greeno (1974) oder aber bei Modellen zum moralischen Urteil nach Kohlberg (Murray et al. (1970)).

#### *Beispiel* (moralisches Urteil)

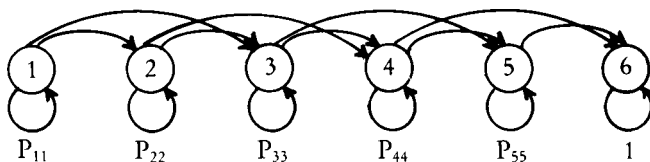
Nach Kohlberg gibt es 6 Stufen der moralischen Entwicklung. Hat jemand eine bestimmte Stufe erreicht, kann er nicht mehr auf eine niedrigere Stufe zurückfallen. Damit ist die oberste Stufe ein absorbierender Zustand. Die Übergangsmatrix von einem Zeitpunkt zum nächsten könnte dann folgende Form haben (s. S. 407 o.).

Man kann die Erreichbarkeit von bestimmten Stufen auch in der Form von gewichteten Graphen darstellen. Dabei treten die Wahrscheinlichkeiten als Gewichte auf (s. Ferschl, 1970; und mit speziellem Bezug zu entwicklungspsychologischen Stufentheorien s. Singer & Spilerman, 1979a).

		t + 1					
		1	2	3	4	5	6
t	Stufe 1	$p_{11}$	$p_{12}$	$p_{13}$	0	0	0
	Stufe 2	0	$p_{22}$	$p_{23}$	$p_{24}$	0	0
	Stufe 3	0	0	$p_{33}$	$p_{34}$	$p_{35}$	0
	Stufe 4	0	0	0	$p_{44}$	$p_{45}$	$p_{46}$
	Stufe 5	0	0	0	0	$p_{55}$	$p_{56}$
	Stufe 6	0	0	0	0	0	1

Dabei wurde noch zusätzlich angenommen, daß in einer Zeitperiode nicht mehr als 2 höhere Stufen erreicht werden können.

Für das Kohlberg-Beispiel ist der Graph:



Es können bei diesen Ketten bezüglich der Zeit bis zur Absorption eine Fülle von Fragestellungen behandelt werden:

- Wie sieht die Verteilung dieser Zeit bei bestimmten Startzuständen aus?
- Wie groß ist die durchschnittliche Zeit bis zur Absorption?
- Wie groß ist die Wahrscheinlichkeit in einem bestimmten Zustand absorbiert zu werden, wenn mehrere Absorptionszustände möglich sind?

Zur Beantwortung solcher klassischer Fragen gibt es eine Reihe von ausgearbeiteten Theorien (siehe Cox & Miller (1968), Ferschl (1970)).

Neben diesen beiden Arten von Ketten gibt es noch weitere Spezialfälle (s. a. Ferschl (1970)).

## 8.2 Markoffketten 2. Ordnung (1 Variable)

Bei Markoffketten 1. Ordnung „beeinflusst“ nur der unmittelbar letzte Zustand vor dem Übergang den Übergang zum nächsten Zustand. Das gilt unabhängig von der Verweilzeit in einem bestimmten Zustand. Falls „mehr Vergangenheit“ berücksichtigt werden soll, muß man zu Ketten höherer Ordnung übergehen. Dabei kann die gewohnte Matrixdarstellung beibehalten werden. Es muß nur die Zahl der Zustände erweitert werden. Das soll an der Erweiterung einer Kette 1. Ordnung zu einer Kette 2. Ordnung erläutert werden

(siehe Anderson (1979)). Markoffketten höherer Ordnung treten speziell bei sozialen Interaktionsprozessen auf (Revenstorf & Vogel, 1979).

Es seien ursprünglich  $K$  Zustände gegeben; so kann man  $K^2$  neue Zustände wie folgt definieren:

t - 2	1 ... 1	2 ... 2	.	..	K...	K	Quellen der Obergänge
t - 1	1 ... K	1 ... K	.	.	.	1 ... K	
neue Zustände	... K K+1 ... 2K ...      ... K <sup>2</sup>						

Das gleiche gilt für die Ziele der Übergänge. Damit geht die Zeitsequenz von je 2 Zeitpunkten in die Definition der neuen Zustände mit ein.

### Fortsetzung des Beispiels:

Gegeben waren jeweils 2 Zustände („gesund“, „krank“). Auf Grund von Tests (siehe später) bzw. theoretischen Überlegungen könnte man zum Schluß kommen, daß für Übergangswahrscheinlichkeiten nicht nur der letzte, sondern die beiden letzten Zustände berücksichtigt werden müssen:

In Tabellenform (Matrixform):

t-2	t-1	1	2
1	1	.90	.10
1	2	.40	.60
2	1	.70	.30
2	2	.20	.80

Bei einer Kette 1. Ordnung müßten die beiden Teiltabellen übereinstimmen.

Falls sie nicht übereinstimmen, ist das ein Indiz für einen Prozeß 2. Ordnung.

Diese Matrixform ist nicht quadratisch wie das etwa bei der Übergangsmatrix  $P$  der Fall war. Das kann man aber leicht durch eine andere Anordnungsart erreichen:

		t - 1	1	1	2	2	
		t	1	2	1	2	
t - 2	t - 1						
1	1		.90	.10	0	0	
1	2		0	0	.40	.60	
2	1		.70	.30	0	0	
2	2		0	0	.20	.80	

$$= \begin{bmatrix} p_{111} & p_{112} & 0 & 0 \\ 0 & 0 & p_{121} & p_{122} \\ p_{211} & p_{212} & 0 & 0 \\ 0 & 0 & p_{221} & p_{222} \end{bmatrix} = P$$

Dabei sind wieder alle Zeilensummen gleich 1. Verschiedene Übergangswahrscheinlichkeiten müssen Null gesetzt werden, da nur jene Übergänge möglich sind, bei denen die Zustände zum Zeitpunkt  $t-1$  übereinstimmen.

Die Verteilungen für bestimmte „Zeitpunkte“ haben dann ebenfalls  $K^2$  Zustände. Die Wahrscheinlichkeiten für Zeitpunkt  $t$  geben an, zum Zeitpunkt  $t-1$  im Zustand  $i$  und zum Zeitpunkt  $t$  in  $j$  zu sein:

$$p'(t) = (p_{11}(t), \dots, p_{ij}(t), \dots, p_{KK}(t))$$

Die Übergangsmatrix  $P$  ist eine  $K^2 \times K^2$  Matrix, deren Elemente zur Hälfte Nullen sind.

Durch diese Umformulierung können die Ketten 2. Ordnung formal gleich behandelt werden wie die Ketten 1. Ordnung. Auch für solche Matrizen kann man den Vektor  $\pi$  berechnen. Er repräsentiert die Wahrscheinlichkeiten, daß jemand bei hinreichend lange dauerndem Prozeß zum Zeitpunkt  $t-1$  in  $i$  und zum Zeitpunkt  $t$  in  $j$  ist.

Weitere Beispiele finden sich bei Anderson (1954, S. 56).

### 8.3 Markoffketten mit mehreren Variablen

Auch die Erweiterung auf mehrere Variable bringt keine grundsätzlichen Schwierigkeiten. Es wird wiederum der Zustandsraum erweitert.

*Beispiel* „Sehen - Kaufen“ nach Anderson (1954)

Die beiden Variablen besitzen folgende Ausprägungen

A: Sehen einer Warenanzeige	ja :	1
	nein :	2
B: Kauf einer Ware	ja :	1
	nein :	2

Für jeden Zeitpunkt gibt es eine Verteilung:  $p'(t) = (p_{11}^{(t)} p_{12}^{(t)} p_{21}^{(t)} p_{22}^{(t)})$ .

wobei  $p_{12}(t)$  bedeutet: Wahrscheinlichkeit, zum Zeitpunkt  $t$  eine Anzeige für die Ware gesehen zu haben ( $A=1$ ) und die Ware nicht gekauft zu haben ( $B=2$ ).

$$p_{12}(t) = W(\{A=1, B=2\})$$

Zudem gibt es die Übergangsmatrix:



$$P = \begin{matrix} & \begin{matrix} 11 & 12 & 21 & 22 \end{matrix} \\ \begin{matrix} 11 \\ 12 \\ 21 \\ 22 \end{matrix} & \begin{bmatrix} p_{11,11} & p_{11,12} & p_{11,21} & p_{11,22} \\ p_{12,11} & p_{12,12} & p_{12,21} & p_{12,22} \\ p_{21,11} & p_{21,21} & p_{21,21} & p_{21,22} \\ p_{22,11} & p_{22,12} & p_{22,21} & p_{22,22} \end{bmatrix} \end{matrix} \quad \text{wobei}$$

$p_{12,21}$  = Wahrscheinlichkeit, zum Zeitpunkt  $t$  eine Anzeige gesehen, aber nicht die Ware gekauft zu haben und danach zum Zeitpunkt  $t + 1$  eine Anzeige nicht gesehen, aber die Ware gekauft zu haben.

$t + 1$

Zu den oben besprochenen Analysen könnte man noch eine weitere anfügen, in der untersucht wird, ob die Veränderungen des *Seh-* und *Kaufverhaltens* unabhängig sind. Bei Unabhängigkeit der Veränderung müßte gelten:

$$p_{ij,hk} = p_{ih}^A \cdot p_{jk}^B$$

Im Beispiel ist:

$$\begin{array}{ll} p_{ih}^A & \text{die Übergangswahrscheinlichkeiten für Sehen (A) und} \\ & \text{die entsprechende Übergangswahrscheinlichkeit für} \\ p_{jk}^B & \text{Kauf (B)} \end{array}$$

Für solche Fragestellungen hat Anderson (1954) ebenfalls einen Test entwickelt.

## 8.4 Schätzung der Übergangswahrscheinlichkeiten

Es sind 3 Fälle zu unterscheiden:

- Stehen sehr lange Beobachtungsreihen ( $T \rightarrow \infty$ ) bei einer Person ( $N = 1$ ) zur Verfügung (Mikrodaten), können die Schätzer von Bartlett (1951) Verwendung finden (siehe auch Anderson & Goodman (1957)).
- Stehen Daten von *vielen* Personen zur Verfügung, bei denen jeweils die vollen Sequenzen bekannt sind, liegen ebenfalls Mikrodaten vor.
- Erhebt man von mehreren Personen nicht unmittelbar die Sequenzen, sondern nur die Häufigkeiten pro Zeitpunkt, liegen Makrodaten vor.

Für die Mikrodaten-Situation haben Anderson & Goodman (1957) Maximum-likelihood-Schätzer entwickelt. Lee, Judge & Zehner (1977) referieren diese und stellen zudem Bayes- und Makrodatenschätzer vor.

Der Einfachheit halber werden wir uns hier mit einigen ML-Schätzern für Mikrodaten begnügen. Hat man folgende Beobachtungen  $y_{vt}$  gemacht, ( $t = 1, \dots, T$  und  $v = 1, \dots, N$ ) können die Werte  $y_{vt}$  selbst jeweils die Nummer eines Zustandes  $i$  ( $i = 1, \dots, K$ ) annehmen.

Aus diesen Sequenzen kann man folgende Häufigkeiten berechnen:

$n_{ij}(t)$  Anzahl der Personen, die zum  
Zeitpunkt  $(t-1)$  in  $i$  und zum  
Zeitpunkt  $t$  in  $j$  sind.

#### 8.4.1 bei Zeitinhomogenität

Es wird nicht vorausgesetzt, daß alle Übergangswahrscheinlichkeiten für die verschiedenen Zeitpunkte gleich sind. Dann erhält man folgende ML-Schätzer:

$$\hat{p}_i(t) = \frac{n_i(t)}{N} \quad i = 1, \dots, K; t = 1, \dots, T$$

$$\hat{p}_{ij}(t-1, t) = \frac{n_{ij}(t)}{n_i(t-1)} \quad i, j = 1, \dots, K; t = 2, \dots, T$$

$$\begin{aligned} \text{Dabei ist } n_i(t-1) &= \sum_{j=1}^K n_{ij}(t) \text{ (Zeilensumme)} \\ &= \sum_{l=1}^K n_{il}(t-1) \text{ (Spaltensumme)} \end{aligned}$$

#### 8.4.2 bei Zeithomogenität

Die Übergangswahrscheinlichkeiten werden als konstant über die Zeitpunkte hinweg vorausgesetzt.

Dann gilt:

$$\begin{aligned} \hat{p}_i &= \frac{n_i}{N} \quad \text{wobei: } n_{ij} = \sum_{t=2}^T n_{ij}(t) \\ \hat{p}_{ij} &= \frac{n_{ij}}{n_i} \quad n_i = \sum_{j=1}^K n_{ij} \end{aligned}$$

Die Schätzer sind approximativ normal verteilt, so daß man leicht Konfidenzintervalle angeben kann. Streuung und Covarianzen der Schätzer finden sich auch bei Anderson (1979).

## 8.5 Tests

Testmöglichkeiten sollten zumindest für einige wichtige Fragestellungen angegeben werden:

a) Spezielle Wahrscheinlichkeiten:

Sind die Wahrscheinlichkeiten gleich speziellen Werten? (etwa in homogenen Ketten)

$$H_0 : p_{ij} = p_{ij}^0; \quad \text{wobei } p_{ij}^0 \text{ ein festgelegter Wert sein kann}$$

b) Zeithomogenität: Sind die Übergangswahrscheinlichkeiten konstant für die verschiedenen Zeitpunkte?

$$H_0 : p_{ij}(t-1, t) = p_{ij}(s-1, s) \quad \begin{matrix} s, t = 2, \dots, T \\ s \neq t \end{matrix}$$

c) Ordnungshypothese: Ist eine Markoffkette eine 1. oder 2. Ordnung?

$$H_0 : p_{ijl}(t-2, t-1, t) = p_{jl}(t-1, t) \text{ für alle passenden } t$$

Diese Fragestellung läßt sich erweitern auf beliebige Ordnungen. Ein wichtiger Spezialfall dabei ist, ob es sich um einen Prozeß 0-ter Ordnung handelt (Unabhängigkeit)

d) Gruppenhomogenität: Gelten für verschiedene Gruppen (z.B. Männer, Frauen) gleiche Übergangswahrscheinlichkeiten?

$$H_0 : p_{ij}^{(m)} = p_{ij}^{(w)} \quad \begin{matrix} m : \text{für „männlich“} \\ w : \text{für „weiblich“} \end{matrix}$$

Für diese und weitere Hypothesen haben Anderson & Goodman (1957) sowohl gewöhnliche Anpassungstests (Pearson) als auch die Likelihood-verhältnistests entwickelt. Wir haben die Teststatistiken für die 4 Arten der Fragestellungen in der Tabelle auf S. 413 zusammengestellt.

## 8.6 Spezielle Probleme und Lösungen bei der Anwendung von Markoffketten

Im Rahmen von wiederholten Messungen an verschiedenen Personen wird vorausgesetzt, daß für jede Person die gleichen Übergangswahrscheinlichkeiten gelten. (Eine Annahme, die auch bei einem gewöhnlichen Regressionsmodell vorliegt, da jeweils unterstellt wird, daß die Regressionskoeffizienten für alle Personen die gleichen sind.)

		$\chi^2$ -Verteilte Teststatistik	Freiheits- grade
Spezielle Wahrschein- lichkeiten	$H_0: p_{ij} = p_{ij}^0$ $j = 1, \dots, K$  $H_0: p_{ij} = p_{ij}^0$ $i = 1, \dots, K$ $j = 1, \dots, K$	$x_i = n_i \sum_{j=1}^K \frac{(\hat{p}_{ij} - p_{ij})^2}{p_{ij}}$  $\sum_{i=1}^K x_i$	$K - 1$  $K(K - 1)$
Zeithomo- genität	$H_0: p_{ij}(t-1, t) =$ $= p_{ij}$ für $i, j = 1, \dots, K$ $t = 2, \dots, T$	$\sum_{t=2}^T \sum_{i=1}^K \sum_{j=1}^K n_i(t-1) \frac{(\hat{p}_{ij}(t-1, t) - \hat{p}_{ij})^2}{\hat{p}_{ij}}$	$K(K - 1)$ $(T - 2)$
Ordnung der Kette (im Bei- spiel: Kette st 1. Ord.) Voraus. : Zeit- homogenität	$H_0: p_{ijl} = p_{jl}$ $i, j, l = 1, \dots, K$	$\sum_{i=1}^K \sum_{j=1}^K \sum_{l=1}^K n_{ij} \frac{(\hat{p}_{ijl} - \hat{p}_{jl})^2}{\hat{p}_{jl}}$	$K(K - 1)^2$
Gruppen- homogenität (unter Vor- aussetzung von Zeithomogen.) g: Anzahl Gruppen	$H_0: p_{ij}^{(h)} = p_{ij}^{(h')}$ $i, j = 1, \dots, K$ $h, h' = 1, \dots, g$	$\sum_{h=1}^g \sum_{i=1}^K \sum_{j=1}^K n_i^{(h)} \frac{(\hat{p}_{ij}^{(h)} - \hat{p}_{ij})^2}{\hat{p}_{ij}}$ wobei $\hat{p}_{ij}^{(h)}$ für die Gruppen ein- zeln, $\hat{p}_{ij}$ für alle Gruppen ge- meinsam geschätzt wird. $n_i^{(h)}$ Anzahl im i ten Zustand in h- ter Gruppe	$(g - 1) \cdot$ $K(K - 1)$

Die Abkürzungen für die Schätzer  $\hat{p}_{ij}$  usw. stimmen mit den oben eingeführten überein.

Bei der Analyse mit Markoffkettenmodellen konnte immer wieder die Beobachtung gemacht werden, daß (8.17) nicht gilt:

$$(8.17) \quad P(1, t) = \prod_{i=2}^t P(i-1, i) \quad \underset{\text{bei Homogenität}}{=} \quad P^{(t-1)}$$

(siehe z.B. Singer & Spilerman\*, 1976; Coleman, 1964). Dabei waren die Diagonalelemente der Matrix  $P^{(t-1)}$  fast immer kleiner als die Elemente der Matrix:  $P(1,t)$ :

$$(8.18) \quad p_{ii}^{(t-1)} < p_{ii}(1,t); \quad p_{ii}^{(t-1)} \in P^{(t-1)}$$

Diese Beobachtung kennen wir aber schon vom obigen Beispiel. Die Gleichheit (8.17) muß ohnehin nur gelten, falls die Markoffeigenschaft gegeben ist. Das hieße dann allerdings, daß in Anwendungen selten Markoffeigenschaft zu erwarten wäre, und daß damit empirische Analysen mit Hilfe von Markoffmodellen wegen nicht erfüllbarer Bedingungen aufgegeben werden müßten.

Dieses Phänomen, so fanden Blumen, Kogan & McCarthy (1955), kann aber auch andere Gründe haben. Es kann sein, daß für zwei Gruppen von Personen jeweils eine Markoffkette (etwa 1. Ordnung) mit nicht identischen Übergangswahrscheinlichkeiten zutrifft. Analysiert man fälschlicherweise beide Gruppen zusammen, betrachtet man eine *Mischung* von Prozessen, für die dann sowohl die etwa vorher vorhandene Homogenität als auch die Markoffeigenschaft nicht mehr gilt. Coleman (1964) wiederum konnte nachweisen, daß dieses Phänomen auch durch *Meßfehler* zustande kommen kann. Das heißt: es gilt zwar für die untersuchte Variable voll die Markoffeigenschaft; da aber die Variable nicht fehlerfrei gemessen werden kann, gilt sie nicht mehr für die meßfehlerbehaftete Variable.

Ein weiterer Grund für die Abweichung von den Markoffannahmen kann auch darin gesehen werden, daß sich die Bedeutung bzw. die Attraktivität der Zustände und damit der Übergangswahrscheinlichkeiten im Laufe der Zeit ändern: bei verschiedenen Berufsgruppen in Mobilitäts-, bei Orten im Rahmen von Wanderungsstudien (Ginsberg, 1971), oder bei sozialpsychologischen Experimenten (Conlisk (1976)).

Für Markoffmodelle gilt, daß die Wahrscheinlichkeit, in einer Klasse länger als eine bestimmte Zeit  $t$  zu bleiben, negativ-exponentiell verteilt ist. Das bedeutet, daß die Wahrscheinlichkeit, sehr lange in einem bestimmten Zustand zu bleiben, mit zunehmender Zeit gegen 0 geht. Dies widerspricht aber vielen in den Sozialwissenschaften betrachteten Prozessen (McGinnis (1968)). Oft trifft zu: „Je länger jemand in einem Zustand war, desto eher bleibt er.“ McGinnis bezeichnete dieses Phänomen als das: „Gesetz der *kumulativen Trägheit*“ (Inertia).

Ein weiteres Problem bei der Anwendung von Markoffketten ist die *Diskrettheit des Zeitparameters*, da ja Übergänge jederzeit stattfinden können (Singer & Spilerman (1976)).

---

\* Singer & Spilerman beziehen sich im Gegensatz zu (8.17) nicht auf die Wahrscheinlichkeiten sondern auf deren Schätzer.

Im Folgenden sollen einige Ansätze besprochen werden, die als Lösungen für die obigen Probleme angesehen werden können.

## 8.7 Einführung unabhängiger Variablen

### 8.7.1 Subgruppenmodelle

Die Gesamtheit wird in mehrere Subgruppen aufgeteilt, für die dann eventuell die Markoffeigenschaften gelten könnten.

### 8.7.2 Übergangswahrscheinlichkeiten als Funktionen von unabhängigen Variablen

Die Aufteilung in mehrere Subgruppen ist in der Praxis nur beschränkt möglich. Je mehr Gruppen es gibt, desto weniger Häufigkeiten stehen für das Schätzen der Wahrscheinlichkeiten zur Verfügung. Deshalb realisierte Spilerman (1972) dann einen Vorschlag von Anderson (1954).

$$(8.19) \quad p_{ij} = \mu_{ij} + b_{ij1} X_1 + \dots + b_{ijm} X_m$$

Die Übergangswahrscheinlichkeiten werden als lineare Funktion eines Globalparameters  $\mu_{ij}$  und von Einflüssen der Eigenschaften der verschiedenen Personen angesehen. Dieses Modell läßt nicht nur zu, die Populationsheterogenität sondern auch das Problem des Attraktivitätswandels der Zustände zu berücksichtigen. Das geschieht, indem in manchen der unabhängigen Variablen die Eigenschaften der Zustände operationalisiert werden\*, (siehe Horan (1976) und Ginsberg (1972)), wobei Ginsberg (1972) allerdings kein lineares Modell sondern ein logistisches\*\* verwendet hat.

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \mu_{ij} + \sum_{e=1}^m b_{ije} X_e$$

Darüber hinaus ist auch die Verwendung von log-linearen oder Probitmodellen möglich.

---

\* Falls in den unabhängigen Variablen nur die Eigenschaften von Zielzuständen (j) operationalisiert werden, die nicht unterschiedliche Effekte für verschiedene Ursprünge (i) sind, kann angenommen werden, daß für alle i die Effekte gleich sind. Dann kann statt  $b_{ij}$ , einfach  $b_j$ , gesetzt werden.

\*\* Ronning (1980) hat für Makrodaten ebenfalls ein logistisches Modell mit den entsprechenden Schätzproblemen vorgestellt.

### 8.7.3 Interaktive Markoffketten

Speziell für Problemstellungen, bei denen das Verhalten, einen Zustand zu wechseln, vom Verhalten anderer Personen abhängt, wurde das „interaktive“ Modell entwickelt (Conlisk (1976)). Conlisk zeigt die Brauchbarkeit des Modells an mehreren Beispielen mit sozialpsychologischen Problemstellungen. Dabei wird angenommen, daß die Wahrscheinlichkeit  $p_{ij}$  zerlegt werden kann nach :

$$p_{ij} = (a_{ij} + b_{ij} * p_i) / S_j \quad S_j \text{ ist eine Normierungsgröße}$$

Die Übergangswahrscheinlichkeit ist einmal bestimmt durch einen konstant bleibenden Übergangsanteil  $a_{ij}$  und durch einen Anteil, der von der Randwahrscheinlichkeit  $p_i$  abhängt. Ist  $b_{ij}$  negativ, wird die Übergangswahrscheinlichkeit kleiner, wenn die Randwahrscheinlichkeit hoch besetzt ist („Exklusivitätswanderung“); ist  $b_{ij}$  positiv, steigt die Wahrscheinlichkeit eines Übergangs jeweils in einen stark besetzten Zustand („Nachahmungswanderung“).

Im wesentlichen ist dieses Modell ein Spezialfall von 8.7.2 wobei hier die Attraktivität bzw. Nichtattraktivität der Zustände durch die Randwahrscheinlichkeiten operationalisiert wird.

## 8.8 Einführung latenter Klassen

### 8.8.1 Mover-Stayer-Modell

Während bei 8.7.1 aufgrund der Kenntnis von Variablenwerten a priori-Gruppen gebildet werden können, für die dann u.U. die Markoffeigenschaft gilt, liegt hier eine anders geartete Situation vor. Es sollen Gruppen a posteriori erschlossen werden. Blumen, Kogan & McCarthy (1955) untersuchten Mobilitätsvorgänge. Sie schlugen für eine adäquatere Beschreibung der Prozesse vor, die Personen in zwei Gruppen zu teilen. Eine Gruppe („stayer“) bleibt tendenziell immer im gleichen Zustand. Ihr Anteil an der Gesamtheit für Zustand  $i$  sei  $S_i$ . Sie besitzt eine besonders einfache Übergangsmatrix: die Einheitsmatrix 1. Die andere Personengruppe („mover“) bewegt sich nach gleichen Übergangswahrscheinlichkeiten von Zeitpunkt zu Zeitpunkt. Ihr Anteil an der Gesamtheit beträgt pro Zustand  $i$  den Wert  $(1 - S_i)$ .

Es läßt sich nicht beobachten, ob eine Person ein Mover oder ein Stayer ist. Diese Eigenschaft muß indirekt erschlossen werden. Es handelt sich hier um ein Modell mit 2 latenten Klassen, für das Goodman (1961) spezielle Schätzmethoden entwickelt hat. Erweiterungen auf mehrere latente Klassen wurden von Spilerman (1972) durchgeführt.

### 8.8.2 Generelles Modell latenter Zustände

Während im Mover-Stayer-Modell die Zustände beobachtbar bleiben, aber 2 latente Gruppen gefunden werden müssen, innerhalb derer dann jeweils ein Markoffprozeß abläuft, werden in diesem Modell die Zustände neu etabliert. Die latenten Zustände sind über die „Responsewahrscheinlichkeiten (RW)“ mit den manifesten (gemessenen) Zuständen verknüpft. Die RW geben an, mit welcher Wahrscheinlichkeit jemand auf einem manifesten Zustand zu finden ist, falls er sich in einem bestimmten latenten Zustand befindet. über diese RW kann man dann versuchen, zu interpretieren, welche Bedeutung den latenten Klassen zukommt. Ähnliche Interpretationsvorgänge finden sich z.B. bei der Faktorenanalyse (mit manifesten Items). Deren Ladungen werden als latente Variablen (Faktoren) interpretiert. Von primärem Interesse sind aber die Übergangswahrscheinlichkeiten zwischen den latenten Klassen.

Eine Fülle solcher Modelle werden in Wiggins (1973) und in Lazarsfeld & Henry (1968) vorgestellt. Die Vielzahl der Modelle ist auf verschiedene Annahmen bezüglich der Randwahrscheinlichkeiten, der latenten Übergangswahrscheinlichkeiten und der Zahl der betrachteten Zeitpunkte zurückzuführen. Für jeden Modelltyp werden ebenfalls die Identifikationsmöglichkeiten untersucht, um Response- und Übergangswahrscheinlichkeiten berechnen zu können. Sind die Parameter identifiziert, können sie mit Methoden von Goodman (1974) geschätzt werden.

Das Modell latenter Zustände kann auch als „Fehlermodell“ interpretiert werden, in dem die latenten Zustände als wahre und die manifesten als die entsprechenden fehlerbehafteten Zustände betrachtet werden. Die Responsewahrscheinlichkeiten geben an, mit welcher Wahrscheinlichkeit jemand in den richtigen oder falschen Zustand klassifiziert wird (vergleiche auch mit dem Alles-oder-nichts-Modell weiter oben). Solche Modelle sind näher bei Murray et al. (1971) dargestellt.

## 8.9 Weitere Modelle: zeitkontinuierliche Markoffprozesse

Falls der Zeitparameter kontinuierlich ist, spricht man von Markoffprozessen (z.B. Cox & Miller, 1968). Man kann die Beschreibung des Prozesses dann nicht mehr wie im diskreten Fall mit einer Differenzengleichung (vgl. 8.5) sondern nur noch mit einer Differentialgleichung vornehmen:

$$(8.19) \quad \frac{dp'(t)}{dt} = p'(t)Q \quad 0 < t < +\infty$$

Die Elemente von  $Q$  werden als Übergangsraten bzw. Intensitäten bezeichnet.



Sie sind keine Wahrscheinlichkeiten. Zwar erhält man als Lösung für die Differenzgleichung (8.5)

$$p'(t) = p'(0) P^t = p'(0) P(0, t) \quad (t = 0, 1, 2, \dots)$$

Diese Gleichung gibt die Zustandswahrscheinlichkeiten zum *diskreten* Zeitpunkt  $t$  in Abhängigkeit vom Anfangszustand  $t=0$  an. Da im diskreten Modell die Zeit „Löcher“ hat, sollte man das kontinuierliche Modell bevorzugen. Die entsprechende Lösung der Differentialgleichung (8.19) ist

$$(8.20) \quad p'(t) = p'(0)e^{Qt}$$

$$\text{wobei gilt:} \quad e^{Qt} = I + tQ + \frac{t^2}{2!} Q^2 + \frac{t^3}{3!} Q^3 + \dots$$

$$q_{ij} > 0 \ (i \neq j) \text{ und } q_{ii} = - \sum_{j=1}^K q_{ij}$$

Sie gibt die Zustandswahrscheinlichkeiten für *jeden beliebigen* Zeitpunkt  $t$  in Abhängigkeit vom Anfangszustand  $t = 0$  an. Identifikation und Schätzung der Intensitäten sind z.Z. noch nicht voll gelöst. Für entsprechende Überlegungen bei nichtprobabilistischen Modellen verweisen wir auf Kapitel 9.

Zeitkontinuierliche Markoffprozesse finden sich meist z.Z. nur in theoretischen Analysen (Barholomew, 1973; Laming, 1973; Schweitzer, 1978). Bei Coleman (1964, 1968, 1980) werden die Übergangsintensitäten  $q_{ij}$  - wie die Übergangswahrscheinlichkeiten in (8.19) - in additive Effekte zerlegt:

$$(8.20) \quad q_{ij} = \mu_{ij} + b_{ij1} X_1 + \dots + b_{ijm} X_m$$

Dabei können die  $x$ -Variablen als die externen Einflüsse auf die Übergangsrate und die Koeffizienten  $b_{ijk}$  als die Stärke dieser Einflüsse interpretiert werden.

Coleman sieht darin die Möglichkeit einer kausalen Interpretation (Coleman (1966)). Einen ähnlichen Weg haben Tuma et al. (1979) eingeschlagen. Andress (1980) bringt Beispiele für praktische Anwendungen.

Semimarkoffprozesse ermöglichen *es, die Zeitpunkte des Wechsels der Zustandes und Übergänge in andere Zustände zu entkoppeln*. Dadurch ist es auch möglich, Konzeptionen wie das „Gesetz der Kumulativen Trägheit“ (siehe oben) mit zu berücksichtigen (siehe z.B.: Ginsberg (1971)). Anwendungen in Entscheidungsprozessen bringt Howard (1971).

## 9. Multivariate „Zeitreiben“- und Panelanalyse mit zeitkontinuierlichen Modellen

### 9.1 „Zeitreihenanalyse“ ( $N=1$ , $T \gg M$ , $M > 1$ )

Bei der Analyse von Zeitreihen kann man auf verschiedene Modellvorstellungen zurückgreifen, die entweder eher *datenorientiert* (wie z.B. die Arima-Modelle) oder eher *theorieorientiert* (wie z.B. die zeitdiskreten oder zeitkontinuierlichen Systemmodelle) sind (s.a. Lewandowski, 1980). Bei der Analyse multivariater Zeitreihen mit rein datenorientierten Modellen (Akaike, 1974, 1976; Hannan, 1970, 1976; Priestley, 1978) sind die Identifikationsschwierigkeiten noch erheblich größer als bei den einfachen Arima-Modellen. Sie dürften noch weiter zunehmen, wenn man zusätzliche Erweiterungen in die Raumdimension vornimmt. Solche Modelle könnten besonders für ökologische Fragestellungen, die Veränderungen von Phänomenen über Zeit und Raum zum Untersuchungsziel haben, von Interesse sein (Pfeifer & Deutsch, 1980).

Wir wollen uns hier mit relativ stark theorieorientierten zeitkontinuierlichen Systemen befassen. Das Interesse an solchen Modellen gründet sich auf vier Überlegungen:

- a) die Zeit hat keine „Löcher“
- b) Prognosen sollen für beliebig wählbare Zeitpunkte möglich sein
- c) Suche nach Parameterinvarianz ist bei Kreuzvalidierungen oberstes Ziel
- d) kausal-inhaltliche Interpretationen sollen auf der Prozeßebene des Phänomens erfolgen

Zu a,b) wäre folgendes zu sagen. Alle zeitdiskreten Modelle sind an einem bestimmten Zeittakt der Messung und Prognose gebunden. Liegen z.B. für den multivariaten autoregressiven Prozeß (7.11) die Meßzeitpunkte  $t$  und  $t-1$  ein Jahr auseinander, sind Prognosen mit (7.13) nur für ganz Vielfache dieser Zeitdistanz möglich. Man kann dann Prognosen nur im Jahresrhythmus erstellen. Für alle anderen Zeitpunkte lassen zeitdiskrete Modelle keine Aussagen zu.

Die Suche nach Parameterinvarianz sollte ein fundamentaler Vorgang bei Kreuzvalidierungen sein. Leider wird er beim Umgang mit zeitdiskreten Modellen nicht berücksichtigt, wenn zwei Forscher für ihre Längsschnittuntersuchung verschiedene Meßintervalle benutzen. Sind die Meßintervalle  $t-(t-1)$  bei zwei Untersuchungen verschieden lang, sind die Parameter der zeitdiskreten Modelle (z.B. die Regressionsgewichte in 7.11) auch dann verschieden, wenn alle Beobachtungen aus einem einzigen zeitkontinuierlichen Modell (z. B. 9.3) stammen! Das bedeutet, daß man Längsschnittuntersuchungen nur

kreuzvalidieren kann, wenn man entweder die Parameter der zeitkontinuierlichen Modelle vergleicht oder in beiden Untersuchungen gleich lange Zeitintervalle zwischen den Messungen vorliegen hat.

Die Überlegung d) bezieht sich auf den wichtigen Aspekt der inhaltlichen Theoriebildung. Man kann an Hand (9.6) und (9.7) zeigen, daß Kausalaussagen im Sinne „Variable i beeinflusst Variable j“ von der Wahl zwischen zeitdiskreter vs. zeitkontinuierlicher Modellbildung abhängen.

Beeinflusst im zeitkontinuierlichen Modell (9.2, 9.3) die Variable i die Veränderung der Variablen j nicht, ist der Parameter  $a_{ji}$  in (9.6) gleich Null. Im entsprechenden zeitdiskreten Modell (9.7) wird aber der entsprechende Regressionsparameter  $f_{ji}$  wegen der Reihenentwicklung (9.8) im allgemeinen ungleich Null sein. Das führt zur Konsequenz, daß man im zeitdiskreten Modell Variableneinflüsse vermutet, die im zeitkontinuierlichen Modell nicht vorhanden sind!

Man muß sich daher bei der Formulierung dynamischer Theorien entscheiden, *auf welcher Zeitebene Formulierung und Hypothesenprüfung* erfolgen sollen. *Man darf zeitkontinuierliche Theorien nicht mit zeitdiskreten Methoden prüfen.*

Bei zeitkontinuierlichen Modellen hegt man die intuitiv einleuchtende Vorstellung von nicht sprunghaften sondern stetigen Variablenänderungen. Sind die im Modell enthaltenen Variablen stetige, nach der Zeit differenzierbare Funktionen, kann man Richtung und Stärke der Variablenänderungen durch deren Ableitungen nach der Zeit angeben (s. z.B. Voedodsky, 1969; McClelland, 1979; Goldstein, 1979).

Jeder, der sich im Rahmen der psychologischen Testtheorie mit Itemcharakteristikkurven befaßt hat, wird als diskretes Modell die Guttman-Skala und als kontinuierliches Modell das Rasch-Modell kennengelernt haben. Neben der üblichen Interpretation als Itemcharakteristikkurven kann man sie aber auch als Entwicklungskurven einer einzigen Person verwenden. Sie beschreiben in beiden Fällen die Lösungswahrscheinlichkeit einer Aufgabe in Abhängigkeit von der Fähigkeit  $\xi$ . Wächst die Fähigkeit einer Person, legt die Itemcharakteristik die erwartete Lösungswahrscheinlichkeit fest. Dabei ist die Form der Kurve für alle Personen gleich, wenn das Modell gilt. So ist die Itemcharakteristik im Birnbaum-Modell (s. Fischer, 1974, S. 204)

$$p(\xi) = \frac{e^{\alpha(\xi - \sigma_j)}}{1 + e^{\alpha(\xi - \sigma_j)}} \quad \sigma_j = \text{Schwierigkeit des Items } j$$

Lösung einer Differentialgleichung

$$\frac{dp(\xi)}{d\xi} = \dot{p}(\xi) = [\alpha(1-p(\xi))] \cdot p(\xi)$$

mit nichtkonstantem Koeffizienten:  $[\alpha(1-p(\xi))]$

Die Veränderung der Lösungswahrscheinlichkeit (bzw. Veränderungsrate) ist eine Funktion der Lösungs- und Versagerwahrscheinlichkeit und einer Konstanten alpha. Ist alpha (= Trennschärfeparameter) gleich 1, liegt der Spezialfall des Birnbaum-Modells nämlich das Rasch-Modell vor. Zusammenfassend läßt sich sagen, daß die Itemcharakteristikkurven in der Testtheorie „Lösungen“ einfacher Differentialgleichungen sind, die das Wachstum von Lösungswahrscheinlichkeiten beschreiben, wenn die Fähigkeit wächst.

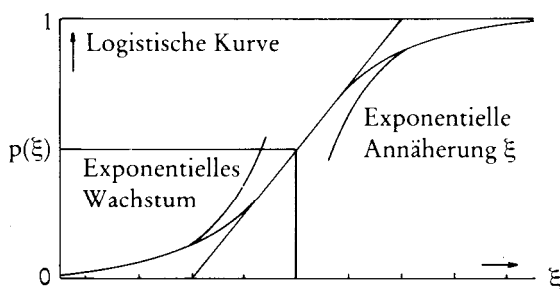


Fig. 9.1: Die logistische Kurve ist die Lösung der Differentialgleichung  $\dot{p}(\xi) = \alpha(1-p(\xi))p(\xi)$ . Sie ist durch drei Parameter bestimmt: den Grenzwert 1, den Mittelwert (Schwierigkeit des Items  $\sigma$ ) und die Ratenkonstante  $\alpha$ , die das exponentielle Wachstum der Anfangsphase und die exponentielle Annäherung an den Grenzwert 1 in der Endphase beschreibt. Dazwischen liegt eine Phase fast linearen Wachstums, die etwa bei  $\sigma-2/\alpha$  beginnt und bei  $\sigma+2/\alpha$  endet. Hier schneidet die Tangente im Mittelpunkt  $\sigma$  die Grenzen 0 und 1 (s.a. Ebenhöf, 1975, S. 44).

Systeme von Differentialgleichungen sind in der Psychologie spätestens seit der Formalisierung der Gruppentheorie von Homans durch Simon (1952, 1957) bekannt geworden. Ausführliche Diskussionen des theoretischen Modells von Simon finden sich bei Rapoport (1963, 1980), Ziegler (1972), Andreski (1977) und Deppe (1977). Nach Deppe läßt sich die Simonsche Formalisierung darstellen als Differentialgleichungssystem (d.h. als gekoppeltes zeitkontinuierliches multivariates System) in zwei Variablen:

1. Gleichung:  $\frac{dS(t)}{dt} = \dot{S}(t) = f_1(A, S)$
2. Gleichung:  $\frac{dA(t)}{dt} = \dot{A}(t) = f_2(A, S, E)$

$$\text{mit: } \frac{\partial f_1}{\partial A} > 0, \frac{\partial f_1}{\partial S} < 0 \quad \text{und} \quad \frac{\partial^2 f_1}{\partial S^2} > 0$$

$$\frac{\partial f_2}{\partial S} > 0, \frac{\partial f_2}{\partial E} > 0, \frac{\partial f_2}{\partial A} < 0 \quad \text{und} \quad \frac{\partial^2 f_2}{\partial A^2} > 0$$

wobei: I = Ausmaß der Interaktion innerhalb einer Gruppe

A = Ausmaß der Aktivität

S = Ausmaß der Sympathie

E = Ausmaß der extern geforderten Aktivität  
(exogene Variable)

Verbal formuliert bedeutet die 1. Gleichung: die Veränderungsrate der Sympathie zum Zeitpunkt t wird gedämpft durch hohe Sympathiewerte und gefördert durch hohe Aktivität. Die 2. Gleichung läßt sich beschreiben als: Die Veränderungsrate der Aktivität zum Zeitpunkt t wird gedämpft durch hohe Aktivitätswerte und gefördert durch hohe Sympathie- sowie extern geforderter Aktivität. Die Lösung des Differentialgleichungssystems würde bei bekannten Werten der Variablen zum Zeitpunkt t=0 Prognosen für jeden beliebigen Zeitpunkt t erlauben, wenn das Modell gilt.

Ähnliche deterministische zeitkontinuierliche Systeme behandelten Rashevsky (1939), Richardson (1948), Blalock (1969), Przeworski & Soares (1977), Wotawa (1979), Rapoport (1980) und im Rahmen der Leistungsmotivationsforschung Loose (1964), Loose & Koran (1975), Loose & Unruh (1977) sowie Kuhl & Blankenship (1979).

Im Gegensatz zu den oben erwähnten Autoren wollen wir uns nicht nur auf theoretische Analysen beschränken, sondern auch die Identifikation und die Schätzung von Systemen (zumindestens ansatzweise) behandeln. Daher beschränken wir uns auf lineare Systeme, deren Koeffizienten nicht von der Zeit abhängen. Die Dynamik eines solchen mehrvariablen Systems kann dann durch eine Anzahl interdependenter Differentialgleichungen und einer Ausgangsgleichung beschrieben werden. Die Gleichungen für den Zeitpfad (Trajektorie) im Variablenraum (Zustandsraum oder „state space“) lauten, wenn die Parameter zeitabhängig sind:

$$(9.1a) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t) \quad \text{Zustandsgleichung}$$

$$(9.1b) \quad y(t) = H(t)x(t) \quad \text{Ausgangsgleichung}$$

mit:  $\dot{x}(t)$  = Vektor mit Ableitungen  $dx_i(t)/dt$  ( $i=1, \dots, M$ )

$A(t), B(t), H(t)$  = zeitabhängige Parametermatrizen, die unbekannt sind

$x(t)$  = Vektor mit Werten der Zustandsvariablen (können auch latente Variablen sein)

$u(t)$  = Vektor mit Werten äußerer Einflußvariabler

$y(t)$  = Vektor mit meßbaren Ausgangsindikatoren des Systems  
zum Zeitpunkt  $t$

Sind, wie wir schon oben angedeutet haben, die Parameter aus Vereinfachungsgründen zeitunabhängig, vereinfacht sich (9.1) zu:

$$(9.2a) \quad \dot{x}(t) = Ax(t) + Bu(t)$$

$$(9.2b) \quad y(t) = Hx(t)$$

Im folgenden wollen wir uns auf das System (9.2) beschränken. Sind alle Variablen  $x$  direkt beobachtbar, kann die Matrix  $H = I$  gesetzt werden. In diesem Fall liegen keine latenten Variablen vor.

$H$  ist daher mit der  $A_y$ -Matrix im Meßmodell von LISREL vergleichbar. Das Prozeßmodell (9.2a) weist aber zum Strukturmodell einen fundamentalen Unterschied auf. In (9.2a) werden die *Veränderungen* von endogenen Variablen beschrieben, während in dem „normalen“ - von Jöreskog bekannten - LISREL-Modell die *Meßwerte* zum Zeitpunkt  $t$  beschrieben werden. Eine dynamisierte Form eines LISREL-Modells wird weiter unten bei den diskreten Approximationen vorgestellt.

Wird  $H = I$  gesetzt, ist die „Lösung“ des linearen Systems (s.a. Chan, Chan & Chan, 1972; Athans, Dertouzos, Spann & Mason, 1974)

$$(9.3) \quad x(t) = e^{A(t-t_0)}x(t_0) + \int_{t_0}^t e^{A(t-\tau)}Bu(\tau)d\tau$$

$$= \left\{ \begin{array}{c} \text{endogene} \\ \text{Entwicklung} \end{array} \right\} + \left\{ \begin{array}{c} \text{exogener} \\ \text{Einfluß} \end{array} \right\}$$

mit der Transitionsmatrix

$$(9.4) \quad e^{A(t)} = I + At + \frac{A^2t^2}{2!} + \frac{A^3t^3}{3!} + \dots = \sum_{K=0}^{\infty} \frac{A^Kt^K}{K!}$$

Die verbale Interpretation der „Lösung“ (9.3) ist einfach. Die Werte des Variablenvektors zum Zeitpunkt  $t$  hängen von den Werten zum Zeitpunkt  $t_0$  (d.h. von der Vorgeschichte), von der endogenen Entwicklung (d.h. der inneren Entwicklung ohne äußere Einflüsse) und dem Einfluß äußerer Variabler ab.

Endogene Entwicklung und äußerer Einfluß überlagern sich bei geeigneter Formulierung des Modells additiv. Die „Lösung“ erlaubt es, bei bekannten Parametermatrizen  $A$ ,  $B$ ,  $H$  und Anfangszustand  $x(t_0)$  und bekanntem Input  $u(\tau)$  (für alle  $\tau < t$ ) den Zustandsvektor  $x(t)$  zu berechnen. Die Beschränkung auf das deterministische System (9.1-9.3) wird weiter unten fallen gelassen.

Werden nur zu diskreten Zeitpunkten  $t=0, T, 2T, \dots$  Beobachtungen erhoben und nimmt man an, daß der Input  $u(t)$  stückweise konstant ist:

$$(9.5) \quad u(\tau): \quad u(kT) = u(kT + \tau) \quad 0 \leq \tau < T$$

So ist es möglich,  $x(t)$  zu den diskreten Zeitpunkten  $t=0, T, 2T \dots$  mittels der diskreten Zustandsgleichung zu bestimmen, wenn wir einmal von Meßfehlern und Prozeßfehlern etc. absehen. Die diskrete Zustandsgleichung lautet:

$$(9.6) \quad x[(k+1)T] = e^{AT}x(kT) + \int_0^T e^{A\tau} d\tau Bu(kT)$$

oder als Differenzengleichung

$$(9.7) \quad \begin{aligned} x[(k+1)T] &= F(T)x(kT) + G(T)u(kT) \\ \text{mit: } F(T) &= e^{AT} \text{ und } G(T) = \int_0^T e^{A\tau} d\tau B = (e^{AT} - I)A^{-1}B = \\ &= A^{-1}(e^{AT} - I)B \end{aligned}$$

zu bestimmen. Die Differenzengleichung (9.7) entspricht genau dem multivariaten autoregressiven Prozeß 1. Ordnung (= multivariate Regression) (7.11) für die Erwartungswerte

$$\begin{aligned} \text{mit } F(T) &= A & \mu(t-1) &= x(kT) \\ G(T) &= b & \text{und } u(kT) &= 1 \end{aligned}$$

(9.7) zeigt ferner, wie das zeitkontinuierliche die Grundlage des zeitdiskreten Modells bildet.

Die Reihenentwicklung von  $F(T)$  kann statt nach (9.4) auch rekursiv erfolgen (Cadzow & Martens, 1970):

$$(9.8) \quad F(T) = e^{AT} \approx \left[ I + AT \left[ I + \frac{AT}{2} \left[ I + \frac{AT}{3} \left[ I + \dots \frac{AT}{L-1} \left[ I + \frac{AT}{L} \right] \dots \right] \right] \right] \right]$$

Für die Matrix  $G(T)$  läßt sich ein ähnlich rekursiver Ausdruck angeben:

$$(9.9) \quad G(T) = (e^{AT} - I)A^{-1}B \approx T \left[ I + \frac{AT}{2} \left[ I + \frac{AT}{3} \left[ I + \dots + \frac{AT}{L-1} \left[ I + \frac{AT}{L} \right] \dots \right] \right] \right] B$$

An Hand der Reihenentwicklung (9.8) wird auch die eingangs entwickelte Überlegung von der Divergenz zeitkontinuierlicher und zeitdiskreter Kausalinterpretationen deutlich. Aus  $a_{ji}=0$  in (9.8) folgt nicht, daß auch  $f_{ji}=0$  ist.

Man kann das inhomogene System (9.2a) in ein homogenes (9.10) umformulieren, wenn  $u(t)$  konstant bleibt. Die Bestimmung von (9.9) erübrigt sich dann:

$$(9.10a) \quad \begin{bmatrix} \dot{x}(t) \\ x(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ B & A \end{bmatrix} \begin{bmatrix} u(t) \\ x(t) \end{bmatrix}$$

$$(9.10b) \quad \dot{x}_e(t) = A_e x_e(t)$$

mit der gegenüber (9.3) vereinfachten Lösung (9.11)

$$(9.11) \quad x_e(t) = e^{A_e(t-t_0)} x(t_0)$$

Nach (9.11) kann man dann (9.7) zur Lösung eines homogenen Systems verwandeln in :

$$(9.12) \quad \begin{bmatrix} u[(k+1)T] \\ x[(k+1)T] \end{bmatrix} = \begin{bmatrix} \\ e^{A_e T} \end{bmatrix} \cdot \begin{bmatrix} u[kT] \\ x[kT] \end{bmatrix} = \begin{bmatrix} I & 0 \\ G(T) & F(T) \end{bmatrix} \cdot \begin{bmatrix} u[kT] \\ x[kT] \end{bmatrix}$$

Wir wollen jetzt ein inhaltliches Beispiel für ein Differentialgleichungssystem (9.2a, 9.3, 9.7) geben. Es findet sich in einer Längsschnittuntersuchung von Doreian & Hummon (1976) zur Analyse der Popularitätszyklen von Regierung, Opposition und Liberalen in England. Als Daten dienten monatliche Umfragen zur Popularität der drei politischen Gruppen und monatlich erhobene Indikatoren wichtiger ökonomischer Variabler (s. Figur 9.2). Das mathe-

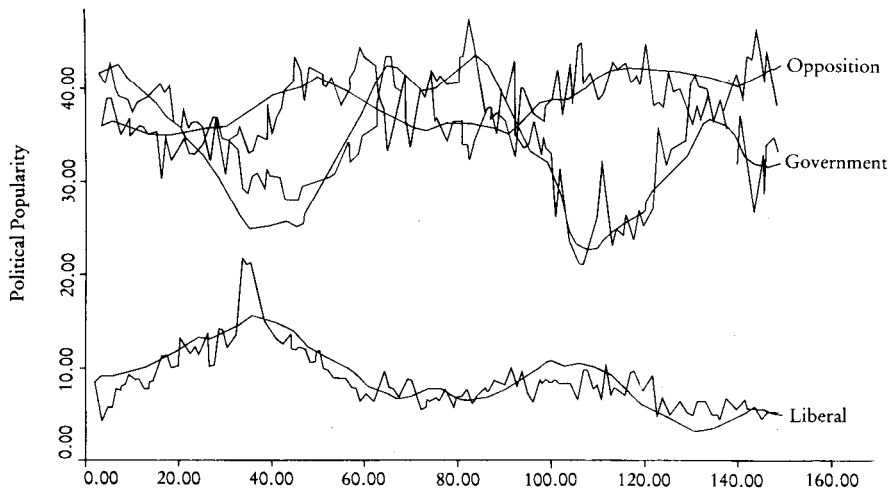


Fig. 9.2: aus Doreian & Hummon (1976): Monatliche Popularitätsentwicklung in England für Regierung, Opposition und Liberale  
Zackenlinie = empirische Zeitreihen  
glatte Kurve = Modellprognosen



Tabelle 9.1: Differentialgleichungsmodell des politischen Klimas in England nach Doreian & Hummon (1976)

$$\begin{bmatrix} \frac{d \text{Pop}_R(t)}{dt} \\ \frac{d \text{Pop}_O(t)}{dt} \\ \frac{d \text{Pop}_L(t)}{dt} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \cdot \begin{bmatrix} \text{Pop}_R(t) \\ \text{Pop}_O(t) \\ \text{Pop}_L(t) \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} & b_{15} & b_{16} \\ b_{21} & b_{22} & b_{23} & b_{24} & b_{25} & b_{26} \\ b_{31} & b_{32} & b_{33} & b_{34} & b_{35} & b_{36} \end{bmatrix} \cdot \begin{bmatrix} \text{In}(t) \\ \text{Un}(t) \\ \text{BP}(t) \\ \text{BR}(t) \\ \text{Cy}(t) \\ C \end{bmatrix}$$

$$\dot{x}(t) = A x(t) + B u(t)$$

$$\begin{aligned} \text{mit: } \frac{d \text{Pop}_i(t)}{dt} &= \text{Veränderung der Popularität von Gruppe } i \begin{cases} i=R \text{ für die Regierung} \\ i=O \text{ für die Opposition} \\ i=L \text{ für die Liberalen} \end{cases} \\ \text{In}(t) &= \text{Inflationsindex (Preisniveau/Lohnniveau) zum Zeitpunkt } t \\ \text{Un}(t) &= \text{Prozentsatz der Arbeitslosen zum Zeitpunkt } t \\ \text{BP}(t) &= \text{Zahlungsbilanz} \left( \frac{\text{Wert der Exporte} - \text{Wert der Importe}}{\text{Preisniveau}} \right) \text{ zum Zeitpunkt } t \\ \text{BR}(t) &= \text{Diskontsatz der Bank von England zum Zeitpunkt } t \\ \text{Cy}(t) &= \text{Scheinvariable, die zwischen zwei Wahlterminen eine umgekehrte U-Funktion beschreibt. Cy soll den Popularitätsbonus der Opposition widerspiegeln. Dieser nimmt in der Mitte zwischen zwei Wahlen für die Opposition ein Maximum an, wie sich an den für England typischen Nachwahlergebnissen ablesen läßt} \\ C &= \text{Scheinvariable, die für alle } t \text{ den Wert } 1 \text{ annimmt} \end{aligned}$$

matische Modell (Tabelle 9.1) repräsentiert die inhaltliche Hypothese, daß Popularitätsänderungen der drei gesellschaftlichen Gruppen sich wechselseitig beeinflussen und von der ökonomischen Entwicklung als äußerem Einfluß abhängen. Das dem System (9.2a) entsprechende inhaltliche Modell findet sich in Tabelle (9.1).

Die Parameter der Systemmatrix  $A$  bestimmen die Entwicklung des Systems vollkommen, wenn keine äußeren Einflüsse vorliegen. Ist z.B.  $a_{12}$  negativ, würde damit die Popularität der Opposition den Zuwachs der Regierungspopularität hemmen. Ist ein Element  $a_{ii}$  auf der Hauptdiagonale von  $A$  negativ, dämpft die Variable  $i$  sich selbst: hohe Werte von Variable  $i$  steuern einen Abfall der Zuwachsrates von  $i$ . So sind z.B. Sättigungs-, Regressions- bzw. Decken(„ceiling“)effekte konsistent beschreibbar.

Kennt man die Paramettermatrizen  $A$  und  $B$ , den Anfangszustand des Variablenvektors  $x(t_0)$  und den Verlauf der äußeren Einflüsse, kann man die Werte der Zustandsvariablen mit der „Lösung“ des Differentialgleichungssystems (9.3) bzw. (9.6, 9.7) prognostizieren. Ideal wäre es natürlich, wenn die äußeren Einflüsse unter vollständiger Kontrolle wären und wenn die Beschreibung des Systems durch  $x$  vollständig wären. Dieser „ideale“ Zustand ist in der Psychologie grundsätzlich nicht erreichbar. Zum einen haben wir es mit offenen Systemen zu tun, die nicht vollständig beschreibbar sind. Zum anderen haben wir das „inverse“ Problem zu lösen: Ausgehend von der Gültigkeit des Modells (9.2) und multivariaten Beobachtungen zu diskreten Zeitpunkten sind die Parameter von  $A$  und  $B$  zu schätzen. Danach erst sind Prognosen mit (9.3) oder (9.7) über die zukünftige Entwicklung von  $x(t)$  möglich. Natürlich hängen diese Prognosen von den Vermutungen oder der Kontrolle über die Rahmenbedingungen (d.h. über  $u(t)$ ) ab.

Das dynamische Verhalten eines Differentialgleichungssystems läßt sich an den Eigenwerten der Matrix  $A$  ablesen. Sind die Eigenwerte von  $A$  verschieden, ist  $A$  kanonisch zerlegbar (Athans et al., 1974):

$$(9.13) \quad A = P \Lambda P^{-1}$$

mit:  $\Lambda$  = Diagonalmatrix mit reellen oder konjugiert komplexen Eigenwerten

$P$  = Matrix, deren Spalten aus den Eigenvektoren von  $A$  bestehen (reell oder konjugiert komplex)

Systeme lassen sich nach ihrem Verhalten in drei Klassen aufteilen (s. Tabelle 9.2; Figur 9.3). So gibt es asymptotisch stabile, stabile und instabile Systeme. Beispiele hierfür finden sich in Figur 9.2. Welcher Fall vorliegt, kann an den Eigenwerten von  $A$  abgelesen werden, ohne für alle  $t \rightarrow \infty$  (9.3) ausrechnen zu müssen. Eine Übersicht über die Kriterien findet sich in Tabelle 9.3.

Tabelle 9.2: Stabilitätsdefinitionen für

asymptotische Stabilität	$\ x(t)\  \rightarrow 0$ bei $t \rightarrow \infty$ (s. Fig. 9.3a)	varianzminimierend
Stabilität	$\ x(t)\  < \infty$ bei $t \rightarrow \infty$ (s. Fig. 9.3b)	varianzminimierend
Instabilität	$\ x(t)\  \rightarrow \infty$ bei $t \rightarrow \infty$ (s. Fig. 9.3c)	varianzmaximierend

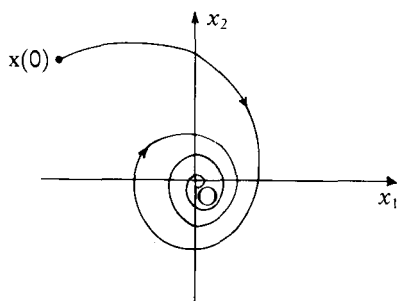
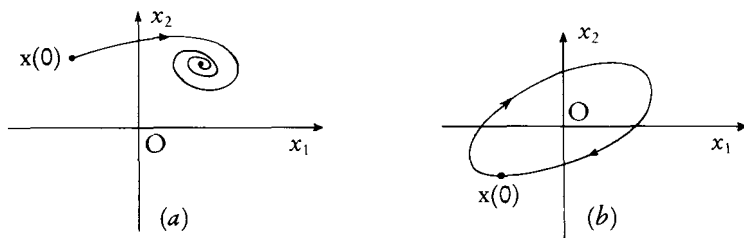
Fig. 9.3a: Asymptotische Stabilität eines Differentialgleichungssystems: dargestellt als Trajektorie (Zeitpfad) im zweidimensionalen Zustandsraum, der durch die Variablen  $x_1$  und  $x_2$  aufgespannt wird

Fig. 9.3b: Stabile Systeme und ihre Trajektorien

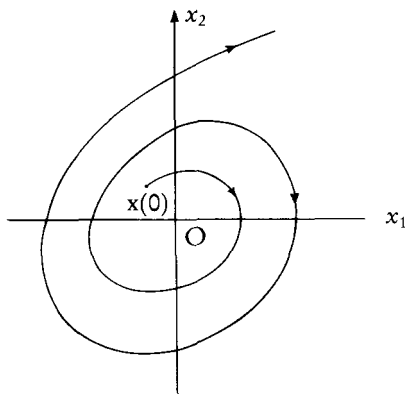


Fig. 9.3c: Beispiel der Trajektorie eines instabilen Systems

Tabelle 9.3: Stabilität und Instabilität von  $\dot{x}(t) = Ax(t)$  in Abhängigkeit der Eigenwerte  $\lambda_i$  der Matrix  $A$ 

asymptotische Stabilität (Figur 9.3a)	alle Eigenwerte von $A$ (einfache oder mehrfache) besitzen negative Realteile $\operatorname{Re}(\lambda_i) < 0$
Stabilität (9.3b)	alle Eigenwerte besitzen nichtpositive Realteile $\operatorname{Re}(\lambda_i) \leq 0$ sind die Realteile $\operatorname{Re}(\lambda_k) = 0$ , müssen die Eigenwerte verschieden sein $\lambda_j \neq \lambda_k$ für alle $j \neq k$
Instabilität (9.3c)	ein oder mehrere Eigenwerte besitzen positive Realteile $\operatorname{Re}(\lambda_i) > 0$ bzw. für einige $j$ und $k$ gilt: $\operatorname{Re}(\lambda_j) = \operatorname{Re}(\lambda_k) = 0$ und $\operatorname{Im}(\lambda_j) = \operatorname{Im}(\lambda_k)$ . Letzteres bedeutet die Existenz mehrfacher Eigenwerte mit verschwindendem Realteil

An dieser Stelle erscheint ein Vergleich mit der Hauptkomponentenanalyse nützlich. Während bei ihr eine *symmetrische* Korrelationsmatrix  $R$  kanonisch nach (9.14) zerlegt wird,

$$(9.14) \quad R = P A P' = P \Lambda P^{-1}$$

um neue *unkorrelierte* Variable zu erhalten, zerlegt man hier die *nichtsymmetrische* Systemmatrix  $A$  nach (9.13), um neue *entkoppelte* Variable zu bestimmen. Während die Eigenwerte in der Hauptkomponentenanalyse die *statischen* Verhältnisse im Sinne von Varianzaufklärungen widerspiegeln, steuern die Eigenwerte des Differentialgleichungssystems die *Wachstumsraten* im neuen entkoppelten Variablensystem. Das kann man daran erkennen, wenn man die Transitionsmatrix  $e^{A(t-t_0)}$  der „Lösung“ (9.3) mit Hilfe von (9.13) anders darstellt:

$$(9.15) \quad e^{At} = P e^{\Lambda t} P^{-1}$$

und

$$e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & 0 & \dots & 0 \\ 0 & e^{\lambda_2 t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & e^{\lambda_m t} \end{bmatrix}$$

$A$  soll verschiedene Eigenwerte besitzen. Sind die Eigenwerte von  $A$  nicht verschieden, kann eine Diagonalisierung nur annähernd über die Jordansche Normalform (Zurmühl, 1968) erfolgen.

### 9.1.1 Stochastische Systeme

Für empirische Fragestellungen in der Psychologie sind die Systeme (9.1) und (9.2) zu idealistisch. Wir müssen von gestörten Variablenbeziehungen und Meßfehlern ausgehen. Ist das stochastische Modell spezifiziert, schließen sich Identifikation und Schätzung des Modells an. Die Schätzung der Parameter wird entweder mittels einer multivariaten Zeitreihe oder mit multivariaten Paneldaten vorgenommen.

Der Einfachheit halber sollen hier nur Systeme 1. Ordnung (d.h. es treten nur einfache Ableitungen im Modell auf) betrachtet werden. Systeme höherer Ordnung können durch geeignete Transformationen auf Systeme 1. Ordnung zurückgeführt werden (s.a. Chan, Chan & Chan, 1972). Ein lineares zeitunabhängiges stochastisches System nimmt dann folgende Form an:

$$(9.16a) \quad \dot{x}(t) = A(\delta)x(t) + B(\delta)u(t) + \varrho(t)$$

$$(9.16b) \quad y(t) = H(\delta)x(t) + \varepsilon(t)$$

wobei:

- $x(t)$  = Vektor mit Zufallsvariablen, die zu diskreten Zeitpunkten beobachtet werden können
- $A(\delta)$  = Parametermatrix, deren Elemente Funktionen eines Vektors  $\delta$  mit Basisparametern sind (im folgenden wird  $A(\delta)$  mit  $A$  abgekürzt)
- $B(\delta)$  = Vektor oder Matrix mit Gewichten für die exogenen Variablen
- $H(\delta)$  = Gewichtsmatrix (Regressionsmatrix) der u.U. latenten Zustandsvariablen  $x(t)$  für die meßbaren Indikatoren  $y(t)$  (entspricht der  $A_y$ -Matrix in LISREL)
- $u(t)$  = Vektor mit exogenen Variablen, die zu diskreten Zeitpunkten beobachtet werden können (muß kontrollierbar sein und darf daher keine Zufallsvariable sein)
- $\varrho(t)$  = Vektor mit Störungen (Fehlern) in den Ableitungen  $\dot{x}(t)$  (= Prozeßfehler)
- $\varepsilon(t)$  = Vektor mit Meßfehlern

Der Vektor  $\varrho(t)$  weicht wesentlich von den normalen bekannten Meßfehlern ab, weil er eine Störung in der Ableitung ist. Er ist daher sehr erratisch. Eine ausführliche Diskussion findet sich z.B. bei Phillips & Wickens (1978, S. 454-458) und Schuss (1980).

Die erratische Charakteristik des Fehlers kann man an einem einfachen Beispiel nur annähernd verdeutlichen. Hat man sich vorgenommen, eine 500 km lange Strecke in 5 Stunden mit dem Pkw zu durchfahren, läßt sich diese (kognitive) Modellvorgabe nicht einhalten. Mai wird die Modellgeschwindigkeit  $dx/dt$  überschritten ( $Q(t) > 0$ ), mal wird sie unterschritten ( $Q(t) < 0$ ),

obwohl am Ende der 5 Stunden das Integral dieses Fehlers sehr klein oder sogar Null sein kann (z.B. nur 1 km bis zum Ziel nach exakt 5 Stunden Fahrt). Weil der Prozeßfehler in stochastischen Differentialgleichungsmodellen grundsätzlich anders ist, als man es aus der „normalen“ Statistik gewohnt ist, machen viele Autoren vereinfachende Annahmen. Oft wird ein deterministisches Prozeßmodell (9.16a) ohne Prozeßfehler mit Meßfehler  $s(t)$  angenommen.

Für die weiteren Betrachtungen sollen die Parameter reell sein. Zusätzlich soll  $A$  verschiedene Eigenwerte mit negativem Realteil besitzen.

Tabelle 9.4: Trivariate Zeitreihe, die dem Modell (9.19) entspricht (aus Phillips, 1972)

$t$	$C(t)$	$Y(t)$	$K(t)$		
0	20.001465	20.001294	40.002760	} $V_{p1}$	Für die im folgenden Abschnitt behandelten Panelmodelle haben wir diese Zeitreihe uminterpretiert: Statt der Zeitreihe einer Person (gemessen an 26 Zeitpunkten) liegen die Daten von 13 Personen (gemessen zu 2 Zeitpunkten) vor.
1	20.724659	21.873653	40.319084		
2	19.500591	20.517932	41.772445		
3	17.740573	16.459774	40.902389	} $V_{p5}$	
4	16.797718	12.794065	36.626464		
5	16.304851	14.501026	32.963211		
6	13.996612	13.712356	32.426635	} $V_{p8}$	
7	15.690959	11.987287	32.672866		
8	13.092237	9.756517	28.897232		
9	14.220060	12.906684	27.340717	} $V_{p11}$	
10	15.704090	16.940200	30.263877		
11	17.718055	17.428257	30.470222		
12	19.487106	23.287891	34.520500	} $V_{p2}$	
13	21.442127	25.636001	37.044532		
14	24.951919	30.484500	45.467407		
15	26.961414	32.732826	48.174156	} $V_{p6}$	
16	27.625804	34.344078	56.707367		
17	29.802745	33.231643	62.022628		
18	30.370506	29.752571	65.508667	} $V_{p9}$	
19	27.389129	21.525676	60.045471		
20	24.136940	18.777145	52.820404		
21	21.597133	17.251140	47.556259	} $V_{p12}$	
22	20.041355	14.616481	42.308754		
23	18.967193	16.165267	35.948036		
24	18.595157	19.830234	32.583084	} $V_{p3}$	
25	20.269153	23.141529	37.360000		

Beobachtungen mit gleichen zeitlichen Abständen, die durch (9.16) erzeugt werden, genügen dem autoregressiven Schema (9.17), wenn  $u(t)$  stückweise konstant ist.

$$(9.17) \quad x(t) = e^A x(t-1) + A^{-1}(e^A - I)Bu(t-1) + \xi(t)$$

$$(9.18a) \quad \text{mit: } \xi(t) = \int_{t-1}^t e^{A(t-\tau)} \varrho(\tau) d\tau$$

und

$$(9.18b) \quad E[\xi(t)\xi(t)'] = \Omega \neq I$$

Die vorstehenden Gedankengänge sollen an einem Beispiel von Phillips (1972) dargestellt werden. Es liegt eine multivariate Zeitreihe (s. Tabelle 9.4) vor, die zu 26 Zeitpunkten „beobachtet“ wurde. Phillips generierte die Zeitreihe nach folgendem stochastischem Differentialgleichungssystem:

$$(9.19) \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 - \lambda & \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \dot{c}(t) \\ \dot{y}(t) \\ \dot{k}(t) \end{bmatrix} = \begin{bmatrix} -\alpha & \alpha(1-\beta) & 0 \\ \lambda & -\lambda & 0 \\ 0 & \gamma v & -\gamma \end{bmatrix} \cdot \begin{bmatrix} c(t) \\ y(t) \\ k(t) \end{bmatrix} + \begin{bmatrix} \alpha \\ 0 \\ 0 \end{bmatrix} [F] + \begin{bmatrix} \varrho_1(t) \\ \varrho_2(t) \\ \varrho_3(t) \end{bmatrix}$$

mit dem Parametervektor  $\delta' = (\alpha, \lambda, \gamma, v, \beta) = (0.6, 4.0, 0.4, 2.0, 0.25)$ , der Matrix  $\Omega \neq I$  und dem konstanten äußeren Einfluß  $F = 5$  für alle Zeitpunkte  $t=0, \dots, 25$ .

Da sich (9.19) nicht in der Form von (9.2a) befindet, muß die zweite Gleichung von (9.19) durch Einsetzen von  $k(t)$  umgeformt werden zu (9.20). Erst (9.20) erlaubt Stabilitätsbetrachtungen.

$$(9.20) \quad \begin{bmatrix} \dot{c}(t) \\ \dot{y}(t) \\ \dot{k}(t) \end{bmatrix} = \begin{bmatrix} -\alpha & \alpha(1-\beta) & 0 \\ \lambda & \lambda(\gamma v - 1) & -\lambda\gamma \\ 0 & \gamma v & -\gamma \end{bmatrix} \cdot \begin{bmatrix} c(t) \\ y(t) \\ k(t) \end{bmatrix} + \begin{bmatrix} \alpha \\ 0 \\ 0 \end{bmatrix} [F] + \begin{bmatrix} \varrho_1(t) \\ \varrho_2(t) \\ \varrho_3(t) \end{bmatrix}$$

Setzt man die Parameterwerte ein, erhält man:

$$(9.21) \quad \begin{bmatrix} \dot{c}(t) \\ \dot{y}(t) \\ \dot{k}(t) \end{bmatrix} = \begin{bmatrix} -.6 & .45 & 0 \\ 4. & -.8 & -1.6 \\ 0 & .8 & -.4 \end{bmatrix} \cdot \begin{bmatrix} c(t) \\ y(t) \\ k(t) \end{bmatrix} + \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} [5] + \begin{bmatrix} \varrho_1(t) \\ \varrho_2(t) \\ \varrho_3(t) \end{bmatrix}$$

mit den Eigenwerten:  $\lambda_1 = -1.56579$

$$\lambda_2 = -0.1171 - 0.37358i$$

$$\lambda_3 = -0.1171 + 0.37358i$$

an den Eigenwerten (vgl. Tabelle 9.3) und den „Daten“ (vgl. Tabelle 9.4) erkennt man die asymptotische Stabilität des Systems.

### 9.1.2 Diskrete Approximation des stochastischen zeitkontinuierlichen Modells

Die Identifikation der Parameter von (9.19) wirft schwierige Probleme auf (vgl. Phillips, 1973).

Zudem setzen Schätzverfahren, die das kontinuierliche Modell direkt schätzen wollen, gute Startwerte voraus, da es sich um ein nichtlineares Schätzproblem (nichtlinear in den Parametern) handelt. Es erhebt sich die Frage, ob es nicht relativ einfache Methoden gibt, mit denen man wenigstens eine diskrete Approximation des kontinuierlichen Modells (9.19) schätzen kann.

Dazu wird jede Einzelgleichung in (9.16) über ein Einheitszeitintervall (Abstand zwischen zwei Messungen)  $(t-1, t)$  integriert. Für die linken Seiten erhalten wir

$$\int_{t-1}^t \dot{x}(\tau) d\tau \approx x(t) - x(t-1) = \Delta x_i(t)$$

Die rechten Seiten enthalten Terme  $\int_{t-1}^t x(\tau) d\tau$ , die nach der Trapezregel durch den Mittelwert  $\bar{x}(t) = \frac{[x(t) + x(t-1)]}{2}$  approximiert werden können. So

wird das Modell (9.19) durch (9.22) diskret approximiert:

$$(9.22) \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 - \lambda & \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \Delta c(t) \\ \Delta y(t) \\ \Delta k(t) \end{bmatrix} = \begin{bmatrix} -\alpha & \alpha(1-\beta) & 0 \\ \lambda & -\lambda & 0 \\ 0 & \gamma v & -\gamma \end{bmatrix} \cdot \begin{bmatrix} \bar{c}(t) \\ \bar{y}(t) \\ \bar{k}(t) \end{bmatrix} + \begin{bmatrix} \alpha \\ 0 \\ 0 \end{bmatrix} \cdot [F] \begin{bmatrix} \eta_1(t) \\ \eta_2(t) \\ \eta_3(t) \end{bmatrix}$$

mit  $\eta(t) = \int_{t-1}^t \varrho(\tau) d\tau$

(9.22) kann dann mit dem bekannten LISREL-Programm von Jöreskog & Sörbom, das ursprünglich für die Schätzung von Strukturgleichungssystemen entwickelt wurde, geschätzt werden. Allerdings sind die Schätzungen, wegen der Linearisierung nicht unverzerrt. Zur Schätzung mit LISREL ändern wir das von Jöreskog (1973) vorgeschlagene Modell\* in einigen Punkten zu folgendem Gleichungstrippel ab:

$$(9.23a) \quad \text{die Struktur- (bzw. hier) Prozeßgleichung} \\ B\eta(t) = \Gamma\xi(t) + \zeta$$

$$(9.23b) \quad \text{das Meßmodell 1: } x(t) = A_y\eta(t)$$

$$(9.23c) \quad \text{das Meßmodell 2: } x(t-1) = A_x\xi(t) = I\xi(t)$$

---

\* Die hier verwendete Symbolik hält sich an Jöreskog; daher ist  $\eta$  in (9.22) von  $\eta$  in (9.23) verschieden.



Tabelle 9.5: LISREL-Modell zur diskreten Approximation des stochastischen Differentialgleichungssystems (9.16) nach dem Vorbild von (9.19)

(9.23a) Strukturgleichung

$\eta_1$

$\Delta C$

$\eta_2$

$\Delta Y$

$\eta_3$

$\Delta K$

$\eta_4$

$\bar{C}(t)$

$\eta_5$

$\bar{Y}(t)$

$\eta_6$

$\bar{K}(t)$

$\eta_7$

$C(t)$

$\eta_8$

$Y(t)$

$\eta_9$

$K(t)$

$\eta_{10}$

$K(t)$

1

0

0

$\alpha$

$-\alpha(1-\beta)$

0

0

0

0

0

0

1

0

0

0

0

0

0

0

0

0

0

1

0

$-\gamma$

$\gamma$

0

0

0

0

0

0

-1

1

0

1

0

0

0

0

1

0

0

-2

0

0

0

0

0

0

0

1

0

0

-2

0

0

0

0

0

0

0

1

0

0

-2

0

0

0

0

1

0

0

0

0

0

0

-1

0

0

0

1

0

0

0

0

0

0

-1

0

0

0

1

0

0

0

0

0

0

-1

$\eta_1$

$\eta_2$

$\eta_3$

$\eta_4$

$\eta_5$

$\eta_6$

$\eta_7$

$\eta_8$

$\eta_9$

$\eta_{10}$

$C(t-1)$

0

0

0

0

0

0

0

0

0

0

0

0

0

-2

0

0

0

0

-1

0

0

0

0

0

0

-2

0

0

0

0

0

0

0

0

-1

0

0

0

0

$\alpha$

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

-1

0

0

0

0

0

0

0

0

0

0

$B\eta = \Gamma\bar{\xi} + \zeta$

$\xi_1$

$\xi_2$

$\xi_3$

$\xi_4$

0

0

0

0

$\xi_1$

$\xi_2$

$\xi_3$

$\xi_4$

0

0

0

0

$C(t-1)$

0

0

0

0

(9.23b) Meßmodell 1

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

0

$x(t) = A_t \eta$

$\eta_1$

$\eta_2$

$\eta_3$

$\vdots$

$\eta_{10}$

$C(t)$

$Y(t)$

$K(t)$

(9.23c) Meßmodell 2

1

0

0

0

0

1

0

0

0

0

1

0

0

0

0

1

$x(t-1) = A_x \bar{\xi}$

$\xi_1$

$\xi_2$

$\xi_3$

$\xi_4$

$C(t-1)$

$Y(t-1)$

$K(t-1)$

$F$

$\eta_7 = \bar{C}(t) + \Delta K - \bar{Y}(t)$

- mit  $A_y$  = Gewichtsmatrix der latenten endogenen Variablen  $\eta$  für die manifesten endogenen Variablen  $x(t)$
- $\eta$  = Vektor mit „latenten“ abhängigen Variablen
- $\xi$  =  $x(t-1)$  = Vektor mit manifesten exogenen Variablen
- $\zeta$  = Vektor mit Gleichungsfehlern (entsprechend dem  $\eta$  in (9.22))

Die Gleichungen (9.23) sind ausführlich in Tabelle 9.5 dargestellt. Die ersten 3 Gleichungen in Tabelle 9.5 bilden das Modell (9.22) ab. Die zweite Gleichung erlaubt es, durch die Einführung einer neuen Variablen  $\eta_7 = \tilde{c}(t) + \Delta K - \bar{y}(t)$  den Parameter  $\lambda$ , der in (9.22) ursprünglich in drei Spalten auftrat, nur einmal zu verwenden. Gleichung 4 in Tabelle 9.5 definiert die Variable  $\eta_7$  aus  $\eta_3, \eta_4, \eta_5$ . Die Gleichungen 5-7 definieren die Variablen  $\eta_4, \eta_5, \eta_6$  (= Mittelwerte) mit Hilfe der  $\eta_1, \eta_2, \eta_3$  (= Differenzen) und der vorherbestimmten  $\xi_1, \xi_2, \xi_3$ . Die  $\eta_1, \eta_2, \eta_3$  (= Differenzen) werden ihrerseits dann in den Gleichungen 8-10 definiert.

Bevor das Modell (Tabelle 9.5) geschätzt werden kann, muß der Identifikationsstatus der Parameter gesichert sein. Parameter sind dann identifiziert, wenn sie eindeutig schätzbar sind. Eine detaillierte Behandlung dieses Problems findet sich bei Phillips & Wickens (1978, S. 451 ff.).

Das Computerprogramm LISREL versucht die aus den Daten (Tabelle 9.4) berechnete empirische Momentenmatrix um Null S (Tabelle 9.6) durch die Matrix 2, die eine Funktion der gesuchten Parameter ist, zu approximieren. Dabei wird die Funktion F (s. Kap. 7.1) minimiert. Eine Gegenüberstellung

Tabelle 9.6: Momentenmatrix S um Null für die Rohdatenmatrix X aus Tabelle 9.4

	C(t)	Y(t)	K(t)	C(t-1)	Y(t-1)	K(t-1)	F
C(t)	445.152						
Y(t)	448.203	464.606					
K(t)	896.869	894.515	1820.809				
C(t-1)	443.137	441.044	897.991	444.721			
Y(t-1)	448.397	456.520	904.930	445.443	459.187		
K(t-1)	890.901	874.339	1815.169	898.583	891.937	1828.987	
F	102.626	101.931	206.544	102.572	101.303	207.073	25.000

$$S = \frac{1}{N} X'X$$

der mit Hilfe von LISREL geschätzten Parameter und der von Phillips vorgegebenen Werte findet sich in Tabelle 9.7. Die durch die diskrete Approximation gewonnenen Schätzungen stimmen für diesen Datensatz gut mit den Parametern überein. Nur  $\hat{\lambda}$  weicht stärker von  $\lambda$  ab.

Tabelle 9.7: LISREL-Schätzungen für die diskrete Approximation eines zeitkontinuierlichen Modells

Parameter	LISREL- Parameter	Startwerte für LISREL	Schätzungen mit LISREL	rückgerechnete Schätzungen für Modell (9.20)
$\alpha = .60$	$\beta_{14} = \alpha$	$= .600$	$\hat{\beta}_{14} = .635$	$\hat{\alpha} = .635$
$\beta = .25$	$\beta_{15} = -\alpha(1-\beta)$	$= .450$	$\hat{\beta}_{15} = -.486$	$\hat{\beta} = .235$
$\lambda = 4.00$	$\beta_{27} = -\lambda$	$= -4.000$	$\hat{\beta}_{27} = -3.201$	$\hat{\lambda} = 3.201$
$v = 2.00$	$\beta_{35} = -\gamma v$	$= -.800$	$\hat{\beta}_{35} = -.819$	$\hat{v} = 2.012$
$\gamma = .40$	$\beta_{36} = \gamma$	$= .400$	$\hat{\beta}_{36} = .407$	$\hat{\gamma} = .407$
	$\gamma_{14} = \alpha$	$= .600$	$\hat{\psi}_{14} = .635$	
	$\psi_{11}$	$= 1.000$	$\hat{\psi}_{11} = 1.321$	Momentenmatrix $\Psi$ der Gleichungs- fehler $\zeta$
	$\psi_{21}$	$= .0$	$\hat{\psi}_{21} = .000$	
	$\psi_{22}$	$= 1.000$	$\hat{\psi}_{22} = 31.200$	
	$\psi_{31}$	$= .0$	$\hat{\psi}_{31} = .000$	
	$\psi_{32}$	$= 1.000$	$\hat{\psi}_{32} = -11.02$	
	$\psi_{33}$	$= 1.000$	$\hat{\psi}_{33} = 4.178$	

Das Anpassungschiquadrat beträgt 4.3896 (df=9) mit einem p=.88, was eine sehr gute Modellanpassung bedeutet. Startwerte mit nur einer Dezimalstelle weisen auf einen konstant gehaltenen Parameter hin. Die rückgerechneten Schätzungen in der rechten Spalte werden durch algebraische Manipulationen aus den LISREL-Schätzungen zurückgerechnet.

9.1.3 Identifikation und Schätzung des zeitkontinuierlichen Systems

Die Parameterschätzung eines zeitkontinuierlichen Modells mit Messungen, die zu diskreten Zeitpunkten erhoben wurden, ist nicht ohne Probleme, wie sich in einer Reihe von Veröffentlichungen gezeigt hat (Coleman, 1968; Land, 1970, 1971; Hummon et al., 1975; Arminger, 1976; Doreian & Hummon, 1976, 1977, 1979). Zum Teil sind die angegebenen „Lösungen“ der Differentialgleichungssysteme nicht korrekt, zum Teil wird dem Fehlermodell zuwenig Aufmerksamkeit geschenkt. Allen Arbeiten aber ist die Nichtbeachtung des Identifikationsproblems gemeinsam: Unter welchen Bedingungen sind die Pa-

parameter aus den Daten eindeutig schätzbar? Zwar sind die Matrizen des zeitdiskreten Systems  $F(T)$  und  $G(T)$  in (9.7) eindeutig und leicht als Regressionsparameter mit LISREL schätzbar. Uns interessieren hier aber die Matrizen des kontinuierlichen Systems  $A$ ,  $B$  und  $H$ . Wie wir weiter unten sehen werden, sind  $A$ ,  $B$  und  $H$  nur unter sehr strengen theoretischen Vorannahmen über die „kausalen“ Beziehungen der Variablen eindeutig aus Fund  $G$  erschließbar. Die Strenge der Vorannahmen geht weit über die Anforderungen bezüglich der Identifizierbarkeit bei normalen Pfad- oder Strukturmodellen hinaus.

Ein Hauptproblem bei der Schätzung liegt in der Mehrdeutigkeit der Matrixgleichung

$$(9.25) \quad F(T) = e^{AT}$$

begründet. Es genügt nicht, wie von den oben genannten Autoren vorgeschlagen, den Logarithmus der Matrix  $F$  zu bilden (Der Logarithmus einer Matrix ist z.B. erklärt bei Gröbner (1966, S. 208)). Der Logarithmus ist nicht definiert, wenn ein Eigenwert von  $F$  im Realteil negativ ist und nicht eindeutig, wenn Eigenwerte konjugiert komplex auftreten, was auf schwingende Zeitreihen hindeutet. Das gleiche Problem tritt auf, wenn man die Intensitäten in zeitkontinuierlichen Markoffprozessen schätzen will (Singer & Spilerman, 1976). Singer & Spilerman (1976) zeigen an einem Beispiel, wie die Transitionsmatrix  $F(T)$  des diskreten Systems vollkommen kompatibel mit zwei Systemmatrizen  $A_1$  und  $A_2$  ist. Jedoch beschreiben  $A_1$  und  $A_2$  zwei vollkommen verschiedene zeitkontinuierliche Veränderungsprozesse:

$$(9.26) \quad F(T) = e^{A_1 T} = e^{A_2 T}$$

$$\text{mit } F(T) = \begin{bmatrix} .234 & .252 & .264 & .250 \\ .252 & .237 & .245 & .266 \\ .268 & .255 & .230 & .247 \\ .248 & .271 & .248 & .233 \end{bmatrix}$$

$$\text{und} \quad A_1 = \quad \quad \quad A_2 =$$

$$\begin{bmatrix} -3.350 & 0.134 & 0.067 & 3.149 \\ 3.132 & -3.306 & 0.144 & 0.030 \\ 0.035 & 3.233 & -3.395 & 0.127 \\ 0.137 & 0.033 & 3.149 & -3.319 \end{bmatrix} \quad \begin{bmatrix} -3.329 & 3.312 & 0.005 & 0.012 \\ 0.033 & -3.337 & 3.209 & 0.095 \\ 0.016 & 0.023 & -3.334 & 3.295 \\ 3.294 & 0.050 & 0.027 & -3.371 \end{bmatrix}$$

Es lassen sich also alternative interne „Kausal“prozesse bei unverändertem Input-Outputverhalten des Modells formulieren.'

Nach  $A_1$  liegt folgende „Kausal“kette vor:

$$X_4 \rightarrow \dot{X}_1, X_1 \rightarrow \dot{X}_2, X_2 \rightarrow \dot{X}_3, X_3 \rightarrow \dot{X}_4$$

und nach  $A_2$  verläuft die Beeinflussung:

$$\dot{X}_4 \rightarrow \dot{X}_3, \dot{X}_3 \rightarrow \dot{X}_2, \dot{X}_2 \rightarrow \dot{X}_1, \dot{X}_1 \rightarrow \dot{X}_4$$

„Many state-variable representations can result in the same input-output system function“ (Athans et al., 1974, S. 430ff.).

Wir wollen nun zeigen, wie Parameter in zeitkontinuierlichen linearen Systemen mit konstanten Koeffizienten identifiziert werden können. Dabei wird sich herausstellen, daß das System (9.20) identifiziert bzw. überidentifiziert und daß das Modell in Tabelle (9.1) *nicht* identifiziert ist. Letzteres trifft auch auf einige Modelle von Coleman (1968), Land (1970, 1971), Hummon et al. (1975) und Arminger (1976) zu, soweit sie als lineare Differentialgleichungssysteme (9.1) oder (9.16) formuliert sind.

Zur Sicherstellung der Identifikation der Parameter müssen den Matrizen Restriktionen auferlegt werden, so daß ausgehend vom Input-Output-Verhalten nur noch eine einzige Modelldarstellung möglich ist.

Das Input-Output-Verhalten eines Systems ohne Berücksichtigung der internen Repräsentation läßt sich durch die Input-Output-Gleichung (s.a. Athans, 1974) beschreiben:

$$\begin{aligned} Y(s) &= G(s) U(s) \\ \text{mit: } Y(s) &= \text{Laplace-transformierte von } Y(t) \\ U(s) &= \text{Laplace-transformierte von } U(t) \\ G(s) &= H(sI - A)^{-1}B = \text{Transfermatrix des transformierten Systems (9.1)} \\ s &= \text{skalare komplexe Variable} \end{aligned}$$

Jedes Element der Transfermatrix steht für eine Input-Output-Verknüpfung. Zum Begriff der Laplace-Transformation sei auf Howard (1971, Vol. II, S. 700ff.), Löhr (1979), McGill (1963) und Restle & Greeno (1970, S. 287f.) verwiesen.

Das Problem der Identifizierbarkeit der Systemparameter reduziert sich dann auf die Frage, ob sich die Parameter in  $A$ ,  $B$  und  $H$  eindeutig aus den Elementen der Transfermatrix  $G(s)$  erschließen lassen (Bellman & Astrom, 1970; Hart & Mulholland, 1979). Dazu muß die Matrix  $G(s)$  symbolisch (d.h. nichtnumerisch) hergeleitet werden. Jedes Element der Transfermatrix  $G(s)$  ist dann ein Bruch von Polynomen in  $s$ . Die Polynomkoeffizienten stellen dann Funktionen der Systemparameter in  $A$ ,  $B$  und  $H$  dar. Diese sind dann identifiziert, wenn bei Kenntnis der Polynomkoeffizienten ein Rückschluß prinzipiell möglich wäre. In diesem Fall würde ein nichtlineares Gleichungssystem mit den im Prinzip aus Messungen bekannten Polynomkoeffizienten  $p$  auf der einen und

den Parametern in A, B und H auf der anderen Seite lösbar oder überbestimmt sein:

$$p = f(A, H, B)$$

Wir wollen dieses an zwei Beispielen demonstrieren. Betrachten wir zunächst das 2-Variablen-Modell:

$$(9.27a) \quad \dot{x}(t) = Ax(t) + bu(t) \quad \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} [u(t)]$$

$$(9.27b) \quad y(t) = Ix(t) \quad \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

Die Inverse  $(sI - A)^{-1}$  läßt sich am einfachsten mit einem Algorithmus von Leverrier (s.a. Frame, 1964; Cadzow & Martens, 1970) bestimmen. Die Transfermatrix  $G(s)$  ist dann ein Vektor mit zwei Elementen, da wir nur eine unabhängige Variable  $u(t)$  und zwei abhängige Variable (Ausgänge)  $y_1(t)$  und  $y_2(t)$  haben. Es gibt daher zwei Eingang-Ausgangsbeziehungen. Für  $G(s)$  finden wir:

$$(9.28) \quad G(s) = (sI - A)^{-1}b = \frac{1}{|sI - A|} \begin{bmatrix} b_1(s - a_{22}) + b_2a_{12} \\ b_1a_{21} + b_2(s - a_{11}) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{sb_1 - (b_1a_{22} - b_2a_{12})}{s^2 - s(a_{11} + a_{22}) - (-a_{11}a_{22} + a_{12}a_{21})} \\ \frac{sb_2 - (b_2a_{11} - b_1a_{21})}{s^2 - s(a_{11} + a_{22}) - (-a_{11}a_{22} + a_{12}a_{21})} \end{bmatrix}$$

Es bleibt zu prüfen, ob die Koeffizienten von  $s$  einen eindeutigen Schluß auf die Modellparameter zulassen. Die Polynomkoeffizienten  $c_j$  lassen sich aus dem Input-Outputverhalten bestimmen und stellen Funktionen der Modellparameter dar:

$$\begin{array}{ll} c_1 = b_1 & c_4 = +a_{11}a_{22} - a_{12}a_{21} \\ c_2 = -b_1a_{22} + b_2a_{12} & c_5 = b_2 \\ c_3 = -a_{11} - a_{22} & c_6 = -b_2a_{11} + b_1a_{21} \end{array}$$

Wir haben jetzt ein nichtlineares Gleichungssystem von sechs Gleichungen in sechs Unbekannten  $(b_1, b_2, a_{11}, a_{12}, a_{21}, a_{22})$ , dessen Lösbarkeit zu prüfen ist.

Betrachten wir nun als zweites Beispiel den Identifikationstatus von (9.20). Phillips (1972) wies in seiner Veröffentlichung ihn nicht nach. Das wollen wir jetzt nachholen. Die Transfermatrix nimmt folgende Gestalt an:

$$(9.29) \quad G(s) = (sI - A)^{-1}B = \frac{1}{|sI - A|} \operatorname{adj}(sI - A)B$$

$$= \begin{bmatrix} \frac{s^2\{\alpha\} + s\{-\alpha\lambda(\gamma v - 1) + \alpha\gamma\} + \alpha\lambda\gamma}{s^3 - s^2\{-\alpha + \lambda(\gamma v - 1) - \gamma\} - s\{\alpha\lambda(\gamma v - \beta) - \gamma(\alpha + \lambda)\} + \alpha\lambda\gamma\beta} \\ \frac{s\{\alpha\lambda\} + \alpha\lambda\gamma}{s^3 - s^2\{.. \quad ..\} - s\{.. \quad ..\} + \alpha\lambda\gamma\beta} \\ \frac{\alpha\lambda\gamma v}{s^3 - s^2\{.. \quad ..\} - s\{.. \quad ..\} + \alpha\lambda\gamma\beta} \end{bmatrix}$$

mit:

$$H = I$$

$$B = \begin{bmatrix} \alpha \\ 0 \\ 0 \end{bmatrix} \quad A = \begin{bmatrix} -\alpha & \alpha(1-\beta) & 0 \\ \lambda & \lambda(\gamma v - 1) & -\lambda\gamma \\ 0 & \gamma v & -\gamma \end{bmatrix}$$

Es ist dann zu prüfen, ob die Koeffizienten der Polynome in  $s$  eindeutig in die Modellparameter  $\alpha, \lambda, \gamma, v, \beta$  rückführbar sind, d.h. das nichtlineare Gleichungssystem (9.30) eindeutig lösbar oder überbestimmt ist:

$$(9.30) \quad \begin{array}{ll} c_1 = \alpha & c_6 = \alpha\lambda\gamma v \\ c_2 = -\alpha\lambda(\gamma v - 1) + \alpha\gamma & c_7 = -(-\alpha + \lambda(\gamma v - 1) - \gamma) \\ c_3 = \alpha\lambda\gamma & c_8 = -(\alpha\lambda(\gamma v - \beta) - \gamma(\alpha + \lambda)) \\ c_4 = \alpha\lambda & c_9 = \alpha\lambda\gamma\beta \\ c_5 = \alpha\lambda\gamma & \end{array}$$

Die Parameter sind identifiziert, weil:

$\alpha = c_1$   $\lambda = c_4/c_1$   $\gamma = c_5/c_4$   $v = c_6/c_5$   $\beta = c_9/c_5$  und überidentifiziert, wenn man noch die anderen Restriktionen hinzuzieht.

Die Identifikation von zeitkontinuierlichen Modellen erweist sich als relativ kompliziert, wenn man mehr als drei Variable betrachtet.

Ist das Modell identifiziert, kann es geschätzt werden. Hierzu bieten sich verschiedene Möglichkeiten an, die von verschiedenen Annahmen bezüglich Fehlerverteilungen, Likelihoodfunktionen und apriori-Verteilungen der Parameter abhängen. So kann man unter der Annahme (9.18b)  $\Omega = I$  eine Datenanpassung nach der Methode der kleinsten Quadrate (OLS) vornehmen:

$$(9.31) \quad \sum \{y(t) - \hat{y}(t)\}' \{y(t) - \hat{y}(t)\} = \min!$$

wobei die Modellprognosen  $\hat{y}(t)$  den Output des Modells (9.16) repräsentieren. Sie werden zweckmäßigerweise über numerische Integration (Shampine, Watts & Davenport, 1976) oder mittels (9.6), (9.7) berechnet. Die Funktion (9.31) kann dann numerisch minimiert werden mit einem Programm zur nichtlinearen Regression (Ralston, Jennrich, Sampson & Uno, 1979).

Eine Schätzung des Systems (9.20) mit den Daten aus Tabelle 9.4 liefert folgende Werte (Tabelle 9.8).

Tabelle 9.8: Direkte Schätzungen eines zeitkontinuierlichen Modells mit der nichtlinearen Regression BMDPAR

Parameter	Startwerte für BMDPAR	Schätzungen mit BMDPAR	Standardschätz- fehler
$\alpha = .60$	.957	.941	.129
$\beta = .25$	.150	.153	.022
$\lambda = 4.00$	1.450	1.275	.383
$\nu = 2.00$	1.979	1.989	.053
$\gamma = .40$	.672	.653	.179
C(0)	} „wahre“ Anfangszustände im Zeitpunkt t=0	16.92	2.624
Y(0)		26.32	2.078
K(0)		44.65	2.188

Fehlerquadratsumme SS = 856.87

Die Schätzungen des hier verwendeten exakten Modells stimmen nicht so gut mit den Parametern überein wie die per LISREL im diskret approximierten Modell gewonnenen (vgl. mit Tabelle 9.7). Jedoch ist der Datenfit relativ gut (s. Figur 9.4), obwohl die Inspektion der Residuen den Verdacht auf positive Autokorrelation aufkommen läßt. Es bleibt zu vermuten, daß das einfache OLS-Kriterium (9.31) wegen (9.18b) nicht angemessen ist.

Für den Fall  $\Omega \neq I$  wurde von Phillips (1972) ein iteratives Schätzverfahren nach der Methode der gewichteten Kleinstquadrate (WLS) vorgeschlagen. Minimiert wird dabei:

$$(9.32) \quad \sum \{y(t) - \hat{y}(t)\}' S \{y(t) - \hat{y}(t)\} = \min!$$

mit positiv definierter Gewichtsmatrix S

Auf die Minimierung der generalisierten Fehlervarianz (= Determinante der Fehlerkreuzproduktmatrix)

$$(9.33) \quad \left| \sum_t \{y(t) - \hat{y}(t)\} \{y(t) - \hat{y}(t)\}' \right| = \min!$$

läuft die Berechnung von Quasi-Maximum-Likelihood (Bergstrom & Wymer, 1976) bzw. Bayes-Schätzern (Box & Tiao, 1973, S. 421ff.) hinaus.

Werden die Modellprognosen Y(t) als „Lösung“ des Differentialgleichungssystems über ein numerisches Integrationsverfahren (z.B. Runge-Kutta) be-



stimmt (s.a. Chan, Chan & Chan, 1972), ist es möglich, mit Dummy- oder Scheinvariablen Interventionseffekte auf der Inputseite des zeitkontinuierlichen multivariaten Systems so zu modellieren, wie es für eine diskrete univariate Zeitreihe von Revenstorff & Keeser (1979, S. 209) und von McCain & McCleary (1979, S. 261 ff.) vorgeführt wurde.

PLOT OF ZEIT VERSUS PREDICTED AND OBSERVED CYK AND VERSUS RESIDUAL

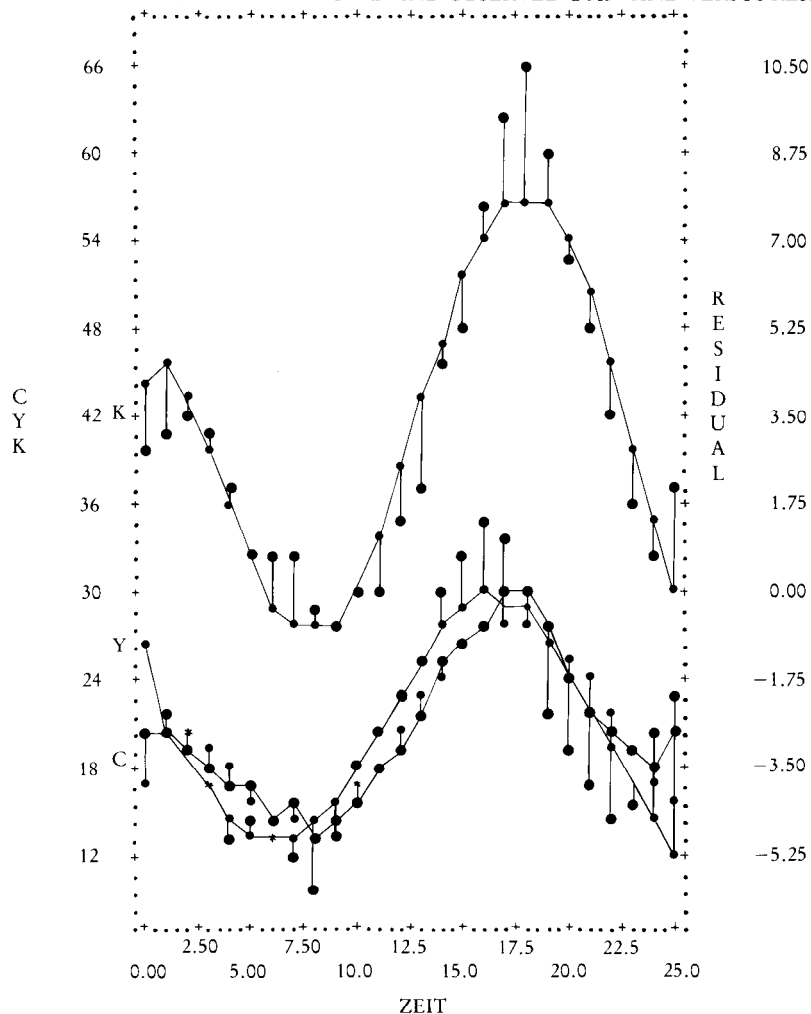


Fig. 9.4: Plot der multivariaten Zeitreihe mit den Variablen  $C(t)$ ,  $Y(t)$ ,  $K(t)$  aus Tabelle 9.4 (schwarze Punkte) und Modellprognose des nach dem Kriterium (9.31) angepaßten Differentialgleichungsmodell (9.20) mit den Parameterschätzungen aus Tabelle 9.8

## 9.2 Panelanalyse (repeated-measurements) ( $N > M$ , $T \geq 2$ , $M > 1$ )

### 9.2.1 Zeitkontinuierliches Modell

Oftmals ist es leichter mehrere Personen zu wenigen Zeitpunkten (z.B.  $T=2$ ) zu beobachten, als an *einer* Person eine ausreichend lange Zeitreihe (z.B.  $T=30$ ). In ersterem Fall liegt dann eine Zeitreihe von Querschnitten (repeated-measurements-design, Panelanalyse) vor. Es ist allgemein üblich, solche Daten varianzanalytisch im Sinne der Wachstumskurvenanalyse zu untersuchen. Dabei werden Entwicklungsverläufe rein deskriptiv durch Polynome beschrieben. Sind  $T$  Zeitpunkte erhoben, lassen sich Polynome bis zum Grade  $T-1$  anpassen. Eine theoretische Begründung für die Beschreibung von Entwicklungsverläufen durch diesen Polynomansatz wird im allgemeinen *nicht gegeben* (s. z.B. Bock, 1979). Außerdem sind bei 2 Messungen nur lineare Extrapolationen als Prognose möglich. Will man mit 2-Punkt-Messungen nicht-lineare Verläufe simulieren oder vorhersagen, muß man systemtheoretische Methoden (s.a. Kap. 7.3 und dieses Kapitel) heranziehen.

Wir setzen hier ein theoretisches Modell für den *Veränderungsprozeß* an. Ähnlich verfahren Singer & Spilerman (1979), jedoch ohne Lösungen für das Identifikations- und Schätzproblem anzugeben.

Ist das Veränderungsmodell formuliert und identifiziert, werden Entwicklungskurven hergeleitet. Diese sind im Gegensatz zum klassischen Wachstumskurvenmodell keine Polynome. Für die Schätzung der Parameter benötigt man mindestens zwei Meßzeitpunkte (Wellen). Ein dritter Zeitpunkt dient zur Testung des Modells. Weicht die Modellprognose (z.B. ein erwarteter Mittelwertsvektor zum Zeitpunkt  $t$ ) signifikant vom empirischen Befund (beobachteter Mittelwertsvektor) ab, gibt es drei mögliche Ursachen:

- a) Das Modell ist falsch spezifiziert. So können die Variablenbeziehungen falsch postuliert sein oder der Prozeß fordert ein Modell mit zeitabhängigen Konstanten (das Modell „altert“).
- b) Wesentliche Variable wurden nicht im Modell aufgenommen (das Modell ist zu einfach, kein empirischer Gehalt)
- c) Die Entwicklung der Rahmenbedingungen  $u(t)$  war entweder nicht unter Kontrolle des Forschers oder wurde von ihm nicht richtig antizipiert. Diese Gefahr ist bei allen Quasiexperimenten und Paneluntersuchungen vorhanden.

Weicht die Modellprognose nicht von den Beobachtungen ab, halten wir das Modell bei. Es zeigt sich hier klarer als bei anderen statistischen Modellen, wie wichtig die äußeren Einflüsse  $u(t)$  für die zeitlichen Verläufe sind. Die Kenntnis der Ausgangswerte  $X(0)$  und die Kenntnis der Modellparameter genügen nicht, um Prognosen richtig zu stellen. Erst wenn die Kontrolle der Rahmenbedingungen  $u(t)$  oder deren Prognostizierbarkeit gewährleistet ist, sind Prognosen möglich.

In einer Test-Retest-Situation (Panelmodell) werden, statt an einem Meßwertträger T Beobachtungen zu erheben, N Personen zu T=2 Zeitpunkten auf M Variablen gemessen. Die „Lösung“ des Systems (9.3) ist in diesem Fall:

(9.34) 
$${}_MX_N(1) = e^A{}_MX_N(0) + \int_0^1 e^{A\tau} d\tau Bu = e^A{}_MX_N(0) + (e^A - I)A^{-1}b$$

wenn die Zeitspanne zwischen den Messungen  $\Delta t = 1$  gesetzt und die äußeren Einflüsse zu einer Variablen  $u(t)$  zusammengefaßt wird, die sich über das Zeitintervall nicht verändert. Zur Verdeutlichung haben wir einmal den Datensatz aus Tabelle 9.4 in einen Datensatz aus einer Paneluntersuchung uminterpretiert und zum anderen drei häufig in der Empirie auftretende Veränderungsprozesse konstruiert. Die Systemgleichungen finden sich in (9.20) und in Tabelle 9.9. Die „Daten“ sind in Tabelle 9.4 und 9.10 aufgeführt.

Der erste Prozeß A weist einen „Regressionseffekt“ auf: Personen mit niedrigen Werten verbessern sich, während sich Personen mit hohen Werten auf beiden Variablen verschlechtern.

Der zweite Prozeß zeigt einen *ceiling- oder Deckeneffekt*: Personen mit hohen Werten verbessern sich kaum noch, während starke Verbesserungen bei Personen mit „Nachholbedarf“ zu beobachten sind.

Der dritte Prozeß C zeigt einen *varianzmaximierenden* Prozeß: die interindividuellen Differenzen werden im multivariaten Raum teilweise größer.

Tabelle 9.9: Koeffizienten der linearen inhomogenen Differentialgleichungssysteme  $X(t) = A x(t-1) + b$  für die Datensätze A, B und C

System A: mit Regres- sionseffekt	$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.39680 & 0.19812 \\ -0.37355 & -0.38008 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 2.13339 \\ 7.84964 \end{bmatrix}$ <p>Eigenwerte: <math>\lambda_1 = -.38844 + .27192 i</math> <math>\lambda_2 = -.38844 - .27192 i</math></p>
System B: mit Decken- od. Sätti- gungseffekt	$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.23180 & -0.13938 \\ -0.09131 & -0.15063 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 6.68670 \\ 4.34701 \end{bmatrix}$ <p>Eigenwerte: <math>\lambda_1 = -.07132</math> <math>\lambda_2 = -.31111</math></p>
System C	$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.18064 & -0.20209 \\ -0.26588 & -0.12308 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 5.39616 \\ 5.51296 \end{bmatrix}$ <p>Eigenwerte: <math>\lambda_1 = -.08172</math> <math>\lambda_2 = -.38544</math></p>

Tabelle 9.10: Fiktive Datensätze A, B und C zur Schätzung von Differenzen- und Differentialgleichungssystemen

Nr. der Person	Datensatz A				Datensatz B				Datensatz C			
	Regressionseffekt				Deckeneffekt							
	Prätests (1. Welle)	Posttests (2. Welle)	Prätests (1. Welle)	Posttests (2. Welle)	Prätests (1. Welle)	Posttests (2. Welle)	Prätests (1. Welle)	Posttests (2. Welle)	Prätests (1. Welle)	Posttests (2. Welle)	Prätests (1. Welle)	Posttests (2. Welle)
	X <sub>10</sub>	X <sub>20</sub>	X <sub>11</sub>	X <sub>21</sub>	X <sub>10</sub>	X <sub>20</sub>	X <sub>11</sub>	X <sub>21</sub>	X <sub>10</sub>	X <sub>20</sub>	X <sub>11</sub>	X <sub>21</sub>
1	1	1	2	5	1	1	6	6	1	1	6	6
2	1	5	3	9	4	7	8	10	7	7	10	10
3	1	8	3	11	1	10	5	13	12	12	13	13
4	1	11	4	13	7	4	11	7	15	15	15,5	15,5
5	1	14	5	15	10	1	14	4	17	17	17,1	17,1
6	2	17	7	16	1	15	4	17	1	4	4	9
7	5	19	8	17	4	12	7,5	14	4,5	10	5,5	13
8	9	20	11	17	7	9	11	10	6	14,5	6,5	16,5
9	12	20	13	17	10	6	14	7	6,5	17,5	6,5	18
10	14	20	14	16	4	16	8	17	6,5	19	6,5	19,1
11	16	19	14	13	1	19	4	19,5	7	2	11	3
12	17	15	14	11	8	18	10	18,5	13	3,5	15,5	4
13	17	10	15	9	7	13	10	14	16,5	4	17,5	4
14	16	7	14	7	15	1	18	1,5	18	4	18,5	4
15	16	4	14	5	18	2	19	4	19	4	19,1	4
16	14	4	12	4	14	6	16	8				
17	8	1	9	5	18	8	19	11				
18	4	1	5	5	13	19	14	19				
19	4	6	6	10	15	19	15,5	19				
20	6	12	8	14	13	15	14	15,5				
21	8	15	11	16	15	14	16	15				
22	13	15	13	12	16	16	16,5	16,5				
23	11	13	12	11	19	13	19	14				
24	9	12	10	13	19	15	19	15,5				
25	10	10	9	11	19	17	19	17,3				
26	14	10	12	9	18	19	18,2	19				
27	13	5	11	7	19	19	19,1	19,1				
28	7	4	8	8	14	12	15	13				

Die Prozeße des „Regressionseffekts“, des „Deckeneffekts“ und der „Differenzierung“ sind in Figur 9.5-7 im bivariaten Testraum durch Veränderungskurven (Zeitpfade, Trajekturen) dargestellt. Punkte stellen die Testwerte dar und Pfeile geben die unter dem jeweiligen Modellparametersatz (Tabelle 9.9) erwarteten Veränderungswege (Lernkurven, Trajektorien) an. Beobachtet

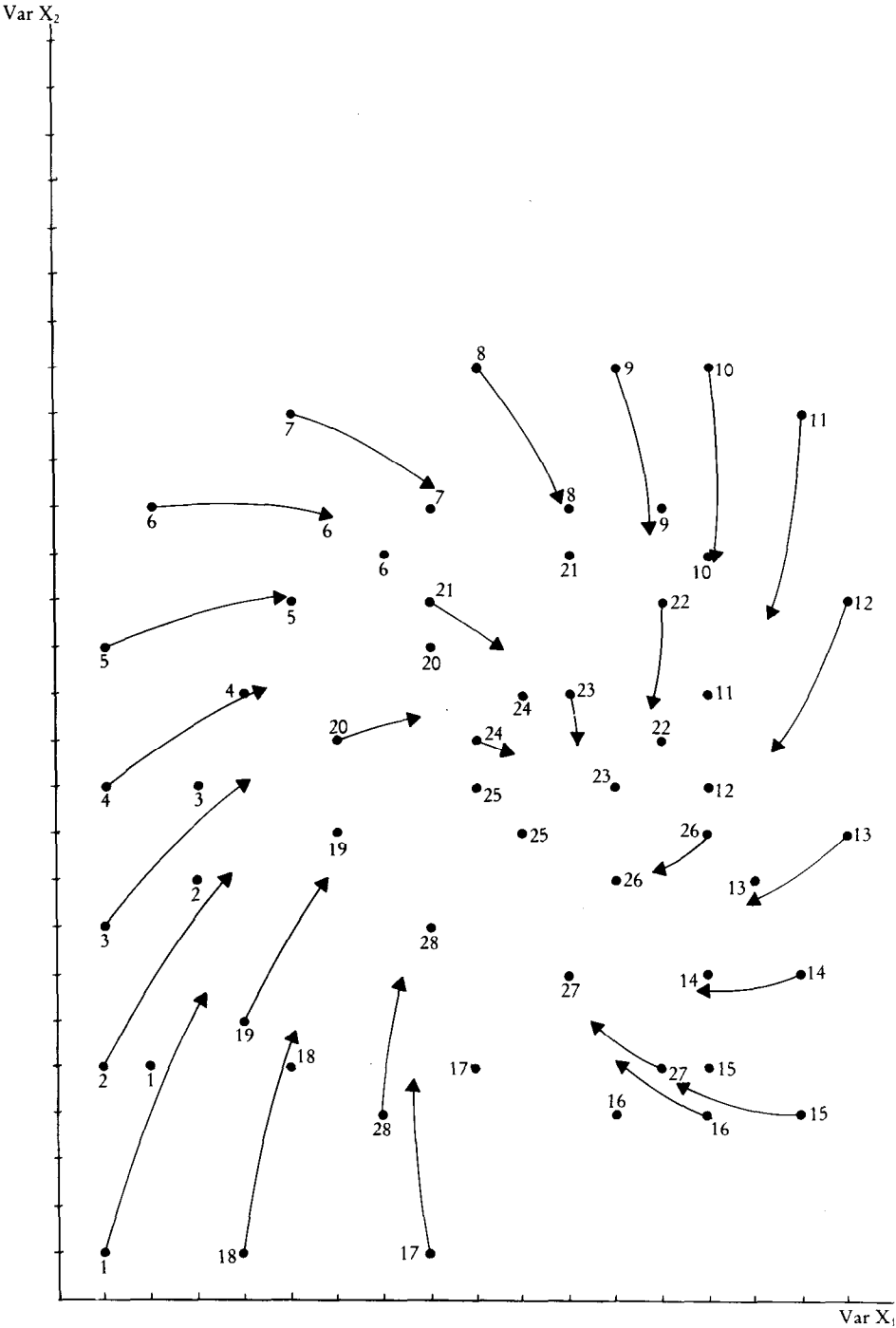


Fig. 9.5: Prozeß A mit „Regressionseffekt“ auf Test X<sub>1</sub> u. X<sub>2</sub>

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.39680 & 0.19812 \\ -0.37355 & -0.38008 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 2.13339 \\ 7.84964 \end{bmatrix}$$

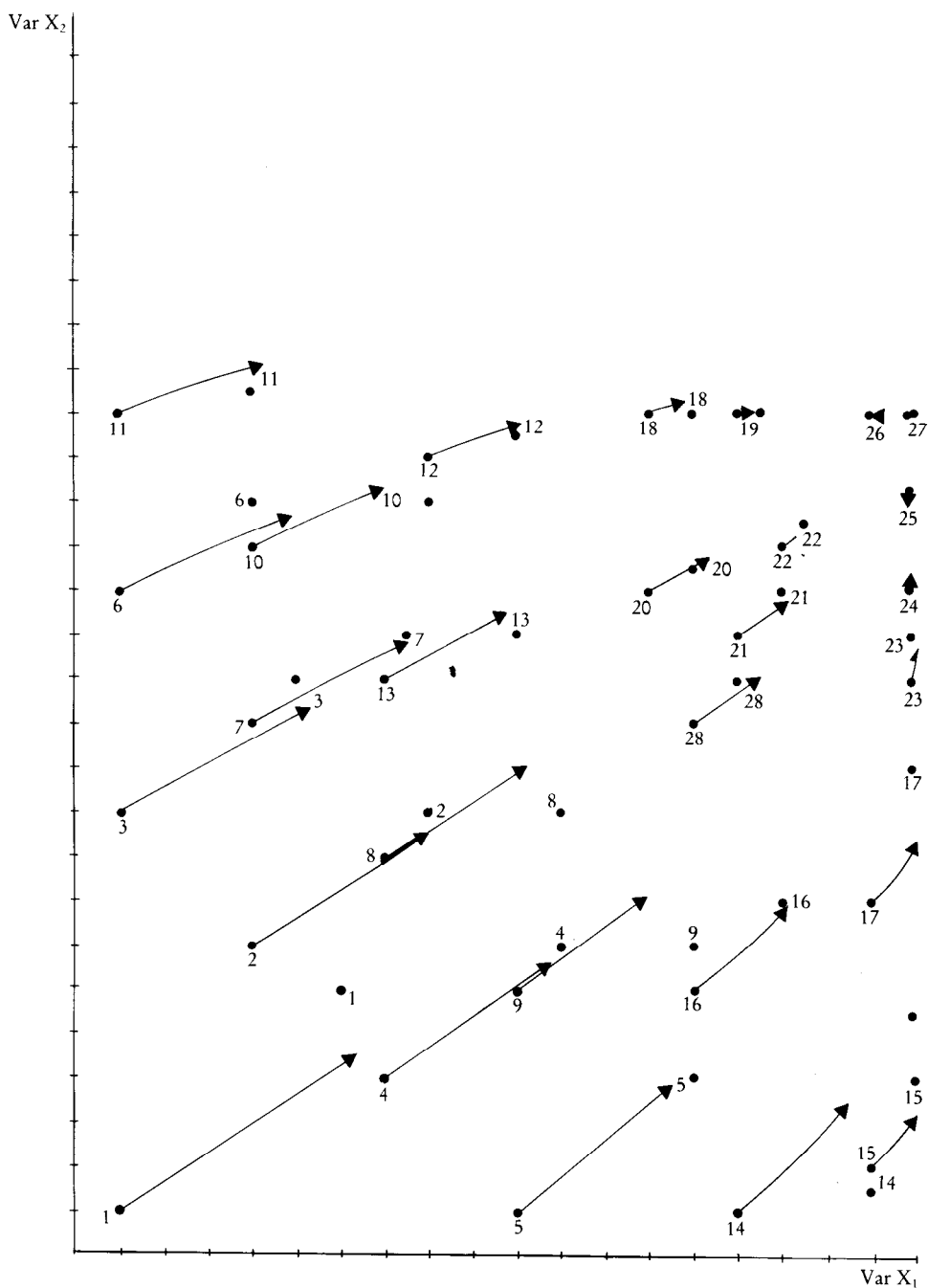


Fig. 9.6: Prozeß B mit „Ceiling- bzw. Deckeneffekt“ auf  $X_1$  u.  $X_2$

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.23180 & -0.13938 \\ -0.09131 & -0.15063 \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 6.68670 \\ 4.34701 \end{bmatrix}$$

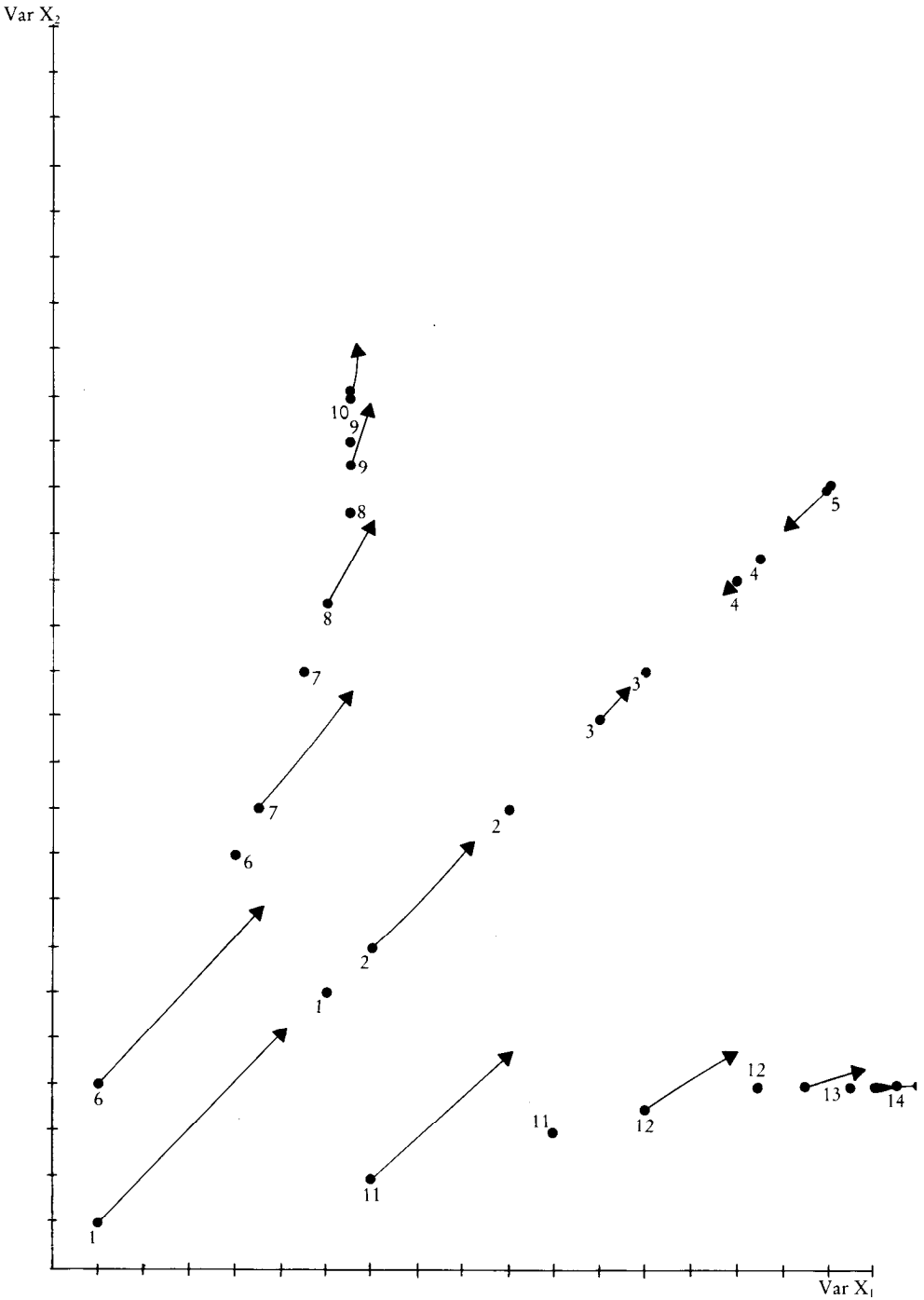


Fig. 9.7: Prozeß C mit „Differenzierungseffekt“, der sich teilweise varianzmaximierend auswirkt

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.18064 & -0.20209 \\ -0.26588 & -0.12308 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 5.39616 \\ 5.51296 \end{bmatrix}$$

werden aber nur Prätest (Anfang des Veränderungsweges) und Posttest. Die erwarteten Veränderungswege gelten unter der Bedingung vorliegender Pretestwerte. Will man zusätzlich „wahre“ Pretestwerte zulassen, müssen längere Zeitreihen vorliegen. Allerdings tritt dann das Problem inzidenteller Parameter auf, weil für jede Person ein derartiger „wahrer“ Pretestvektor eingeführt wird (s.a. Fischer, 1974, S. 350f.; Neyman & Scott, 1948).

Für alle Überlegungen gilt hier folgende Rahmenbedingung: während alle Parameter für alle Personen gleich sind, dürfen Anfangswerte (Pretests), Endwerte (Posttests) und damit die erwarteten Verläufe (Lern- oder Entwicklungskurven) von Person zu Person verschieden sein. *Damit ist es möglich, die verschiedensten Effekte (Regressionseffekt, Ceilingeffekt etc.) unter einem einheitlichen Prozeßgesichtspunkt zusammenzufassen.*

### 9.2.2 Diskrete Approximation des stochastischen zeitkontinuierlichen Panelmodells mit LISREL

Die Approximation der Systeme A („Regressionseffekt“), B („Ceiling-Effekt“) und C (Tabelle 9.9) mittels LISREL erfolgt genauso wie es für die multivariate Zeitreihe im Abschnitt 9.1.2 gezeigt wurde. Jedoch muß das in Tabelle 9.5 gezeigte Modell an einigen Punkten abgeändert werden. Es liegen ja nur zwei Variable vor und die Matrix A weist keine Nulleinträge auf. Die Ergebnisse der linearen Approximation finden sich in Tabelle 9.11.

Tabelle 9.11: Approximation der Systeme A („Regressionseffekt“), B („Ceilingeffekt“) und C („Differenzierung“) mit LISREL

System A: „Regressionseffekt“

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.399 & +0.192 \\ -0.362 & -0.382 \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 2.217 \\ 7.750 \end{bmatrix}$$

Matrix  $\Psi = \begin{bmatrix} 0.756 & \\ 0.403 & 1.245 \end{bmatrix}$  alle | t-Werte | > 1.976  
p=1.00 mit df=0

System B: „Ceilingeffekt“

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.230 & -0.138 \\ -0.090 & -0.150 \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 6.631 \\ 4.315 \end{bmatrix}$$

Matrix  $\Psi = \begin{bmatrix} 0.224 & \\ -0.129 & 0.552 \end{bmatrix}$  alle | t-Werte | > 1.778  
p=1.00 mit df=0

System C: „Differenzierung“

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.178 & -0.197 \\ -0.260 & -0.121 \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 5.301 \\ 5.412 \end{bmatrix}$$

Matrix  $\Psi = \begin{bmatrix} 0.637 & \\ 0.000 & 0.792 \end{bmatrix}$  alle | t-Werte | > 2.63  
p=0.2995 mit df=1



Die angeführten Modelle  $\dot{x}(t) = Ax(t) + b$  beziehen ihre Parameter aus einem LISREL-Modell, das ähnlich zu dem in Tabelle 9.5 gezeigtem ist. Dabei bildet die Matrix  $-A$  einen Teil der Betamatrix und der Vektor  $b$  einen Teil der  $\Gamma$ -Matrix in LISREL. Bei Modellen mit mehr Variablen entscheidet die Auswahl der Startwerte über den Erfolg des Schätzverfahrens. Eine Prozedur, die sich für die Startwertbestimmung bis zu 9 Gleichungen bewährt hat ist in Möbus (1983) beschrieben.

---

Obwohl die über LISREL gewonnenen Schätzungen nicht mehr als 0.1 von den Parametern in Tabelle 9.9 abweichen, sollte vor einer Generalisation dieser erfreulichen Ergebnisse auf „echte“ Daten gewarnt werden. Die Güte der Approximation dürfte von Faktoren wie Fehlervarianzen, Schnelligkeit der Veränderungen und Zeitintervall zwischen den Messungen abhängen. Es werden hierzu noch genauere Studien notwendig sein.

### 9.2.3 Identifikation und Schätzung der zeitkontinuierlichen Panelmodelle

Lassen sich nicht alle Parameter eindeutig schätzen, oder müssen aus inhaltlichen Gründen Modelle restringiert werden, können Parameter auf Konstante fixiert werden oder es können Gleichheitsrelationen zwischen den Parametern spezifiziert werden (so z.B. für d. „Regressionseffekt“). Eine solche Annahme wäre z.B.  $a_{11} = a_{22}$ :

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{11} \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} [u(t)]$$

Nach (9.28) vereinfacht sich das nichtlineare Gleichungssystem zur Identifikationsprüfung zu

$$\begin{array}{ll} c_1 = b_1 & c_4 = -(a_{11}^2 + a_{12}a_{21}) \\ c_2 = -(b_1a_{11} + b_2a_{12}) & c_5 = b_2 \\ c_3 = 2a_{11} & c_6 = -(b_2a_{11} + b_1a_{21}) \end{array}$$

Damit sind die Parameter identifiziert, wie sich durch Einsetzen feststellen läßt:

$$\begin{array}{ll} b_1 = c_1 & a_{21} = (-c_6 - c_5c_3/2)/c_1 \\ b_2 = c_5 & a_{12} = (-c_2 - c_1c_3/2)/c_5 \\ a_{11} = c_3/2 & \end{array}$$

Schätzungen mit dem Programm BMDPAR (Möbus, 1983) nach dem einfachen OLS-Kriterium (9.31) lauten für das nichteingeschränkte Modell mit  $a_{11} \neq a_{22}$ :

Tabelle 9.12: Direkte Schätzung des Systems A „Regressionseffekt“ mit BMDPAR und numerischer Bestimmung der Veränderungsverläufe (= Lernkurven, Trajektorien) über numerische Integration des Differentialgleichungssystems

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -.39671 & +.19816 \\ -.37351 & -.38007 \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 2.13292 \\ 7.84969 \end{bmatrix}$$

Fehlerquadratsumme: 37.72

An den negativen Koeffizienten  $a_{11}$  und  $a_{22}$  erkennt man die Selbstdämpfung der Variablen: Je höher die Pretestwerte  $X(0)$  einer Person, desto kleiner der Zuwachs. Damit ist ein typischer Regressionseffekt auf beiden Variablen  $X_1$  und  $X_2$  beschrieben. Es sollten in Zukunft individuelle Veränderungsprozesse nicht mehr durch die Differenz  $x_i(t) - x_i(0)$  der Meßwertvektoren der Person  $i$ , sondern durch deren Pretestvektor  $x_i(0)$ , dem Parametersatz A, B **und** dem Verlauf der äußeren Einflüsse  $u(t)$  beschrieben werden. Der Einfachheit halber, haben wir hier angenommen, daß sich alle äußeren Einflüsse in einer Summationsvariablen  $u(t)$  vereinigen ließen. Dabei sollten sich die Stärke der Variable  $u(t)$  in der Zeit zwischen Post- und Prätest nicht ändern:  $u(t) = 1$ . Diese Einschränkung kann man im Prinzip fallen lassen, wenn man  $u(t)$  öfter mißt oder die Entwicklung von  $u(t)$  selbst steuert oder den Verlauf von  $u(t)$  prognostizieren kann. Steuert der Forscher die unabhängige Variable  $u(t)$ , liegt eine dynamische Version des klassischen Experiments vor. Dazu gibt es eine sehr lesenswerte Arbeit von Thalmaier (1979), die sich mit der kognitiven Bewältigung der optimalen Steuerung eines dynamischen Systems befaßt.

Die Systeme B („Ceilingeffekt“) und C („Differenzierung“) lassen sich ähnlich identifizieren und schätzen. Bei dem Modell (9.20) wurde die Identifizierung der Parameter schon in (9.29) und (9.30) gezeigt. Um jetzt den ursprünglichen

Tabelle 9.13 : Direkte Schätzung des Modells (9.20) mit Paneldaten aus Tabelle 9.4 mittels BMDPAR und numerischer Bestimmung der Veränderungsverläufe

Parameter aus Modell (9.20)	Schätzungen mit nichtlinearer Regression BMDPAR
$\alpha = .600$	.620
$\beta = .250$	.198
$\lambda = 4.000$	3.231
$\nu = 2.000$	1.994
$\gamma = .400$	.384

als Zeitreihe angelegten Datensatz in einen Datensatz aus einer Test-Retest- (oder Panel-)untersuchung umzuinterpretieren, haben wir jeweils zwei Zeitpunkte zum Datensatz einer Person zusammengefaßt. Es ergaben sich also 13 „Versuchspersonen“. Die Daten wurden dann entsprechend der Personennummer (rechter Rand der Tabelle 9.4) in eine neue Rangreihe gebracht (gemischt), um auch den letzten Eindruck einer Zeitreihe zu verwischen. Die Schätzungen nach BMDPAR (Fortran-Code in Möbus, 1983) sind in Tabelle 9.13 aufgeführt.

Dabei stimmen die Schätzungen mit Panel„daten“ (Tab. 9.13) besser mit den wahren Parametern  $\alpha$ ,  $\beta$ ,  $\lambda$ ,  $\gamma$ ,  $\nu$  überein als die Schätzungen auf der Basis der Originalzeitreihe (Tab. 9.8)! Dieses ist ein weiterer Hinweis auf die Autokorreliertheit der Residuen.

## 10. *Schlußbemerkungen*

Wir haben die Veränderungsmessung, -analyse und -prognose unter dem Blickwinkel betrachtet, daß mit den Ergebnissen einer empirischen Längs-Schnittuntersuchung prinzipiell Handlungsanweisungen zur Optimierung von psychologischen Interventionen gegeben werden können.

Für dieses Ziel sind einige Methoden (z.B. Arima(p,d,q)-Modelle, varianzanalytische Wachstumskurvenmodelle, einfache Markoffketten, Pfadanalysen mit Korrelationsmatrizen) weniger und andere (Transfermodelle, Systemtheoretische Modelle) besser geeignet.

Für alle Typen von Anwendungen gilt aber, daß Veränderungsmessung nicht bei einer individuellen Z-Punkt-Messung und deren Interpretation („Person A hat sich von  $t_1$  nach  $t_2$  um  $\Delta X$  verbessert“) stehen bleiben darf, sondern, daß die Messungen als eventuell meßfehlerbehaftet und als Realisationen eines multivariaten stochastischen Prozesses, der äußeren Einflüssen unterliegt, begriffen werden müssen. Veränderungsmessung bedeutet dann:

- a) Quantifizierung der Realisationen  $x_i(t)$  des Pbn i
- b) Schätzung der Prozeßparameter
- c) Quantifizierung bzw. Kontrolle der exogenen Einflüsse

Bei der Einzelfallanalyse benötigt man hierzu längere Zeitreihen. Die Zahl der Zeitpunkte kann reduziert werden, wenn man mehrere Personen simultan betrachtet und die Annahme aufstellt, daß alle Personen mit Prozessen beschreibbar sind, die gleiche Parameter besitzen (z.B. Panelanalyse).

Legt man diese Maßstäbe an, kann man eine Vielzahl von Methoden als wenig richtungsweisend einstufen und somit ein wenig Ordnung in das Methoden-„chaos“ bringen.

## Literatur

- Aitken, A. C. 1935. On Least Squares and Linear Combination of Observations, *Proceedings of the Royal Society of Edinburg*, 55.
- Akaike, H. 1974. Markovian Representation of Stochastic Processes and its Application to the Analyses of Autoregressive Moving Average Processes. *Annals of the Institute of Statistical Mathematics*, 26, 363-387.
- Akaike, H. 1976. Canonical Correlations Analysis of Time Series and the Use of an Information Criterion. In: Mehra, R. & Lainiotis D. G. (Eds.): *Advances and Case Studies in System Identification*. New York: Academic Press.
- Algina, J. & Swaminathan, H. 1977. A Procedure for the Analysis of Time-Series Designs. *J. of Experimental Education*, 45, 56-60.
- Algina, J. & Swaminathan, H. 1979a. Application of Growth Curve Methodology to the Analysis of Interrupted Time Series Designs. Paper presented at the AERA-Congress, San Francisco.
- Algina, J. & Swaminathan, H. 1979b. Alternatives To Simonton's Analyses of the Interrupted and Multiple-Group Time-Series Designs. *Psychological Bulletin*, 86, 919-926.
- Amemiya, T. 1967. A Note on the Estimation of Balestra-Nerlove Models. Technical Report Nr. 4., Inst. of Math. Studies in the Social Sciences. Stanford University.
- Anderson, O. D. 1975. *Time Series Analysis and Forecasting: The Box-Jenkins Approach*. London: Butterworths.
- Anderson, T. W. 1954. Probability Models of Analyzing Time Changes in Attitudes. In: Lazarsfeld, P. F. (Ed.): *Mathematical Thinking in the Social Sciences*. Glencoe.
- Anderson, T. W. 1971. *The Statistical Analysis of Time Series*. New York: Wiley.
- Anderson, T. W. 1978. Repeated Measurements in Autoregressive Processes. *Journal of the American Statistical Association*, 73, 371-378.
- Anderson, T. W. 1979. Panels and Time Series Analysis: Markov Chains and Autoregressive Processes. In: Merton et al. (Eds.): *Qualitative and Quantitative Social Research*. New York.
- Anderson, T. W. & Goodman, L. A. 1957. Statistical Inference about Markov Chains. *Annals of Math. Statistics*, 28, 89-110.
- Andreski, S. 1977. *Die Hexenmeister der Sozialwissenschaften*. München: Deutscher Taschenbuch Verlag.
- Andress, H. J. 1980. Methoden temporaler Analyse. Arbeitsbericht Nr. 9, Universität Bielefeld, Fakultät für Soziologie.
- Arminger, G. 1976. Analyse und Auswertung von Paneluntersuchungen. In: Holm K. (Hrsg.): *Die Befragung*, 4, 134-235. München: Francke.
- Athans, M., Dertouzos, M. L., Spann, R. N. & Mason, S. J. 1974. *Systems, Networks and Computation: Multivariable Methods*. New York: Mc Graw Hill.

- Atkinson, R. C. & Estes, W. K. 1963. Stimulus Sampling Theory. In: Luce, R. D., Bush, R. R., Galanter, E. (Eds.): *Handbook of Mathematical Psychology* (Vol. II), 121-268.
- Balestra, P. & Nerlove, M. 1966. Pooling Cross Section and Time Series Data in the Estimation of a Dynamit Model: The Demand for Natural Pass. *Econometrica* (34), 585-612.
- Baltes, P. B. 1979. Einige Beobachtungen und Überlegungen zur Verknüpfung von Geschichte und Theorie der Entwicklungspsychologie der Lebensspanne. In: Baltes, P. B. & Eckensberger, L. H. (Hrsg.): *Entwicklungspsychologie der Lebensspanne*. Stuttgart, 13-33.
- Baltes, P. B. & Nesselroade, J. 1979. Die entwicklungspsychologische Analyse von individuellen Unterschieden in mehreren Meßgrößen. In: Baltes, P. B. & Eckensberger, L. H. (Hrsg.): *Entwicklungspsychologie der Lebensspanne*. Stuttgart, 145-178.
- Bartholomew, D. J. 1967. *Stochastic Models for Social Processes*. New York: Wiley.
- Barlow, D. H., Hersen, M. 1973. Single-Case Experimental Designs. *Archive of General Psychiatry*, 29, 319-325.
- Designs für Einzelfalexperimente. In: Petermann, F. (Hrsg.): 1977, *Methodische Grundlagen klinischer Psychologie*. Weinheim: Beltz, 64-83.
- Bartlett, M. S. 1947. Multivariate Analysis. *Journal of the Royal Statistical Society Supplement, Series B*, (9), 176-197.
- Bartlett, M. S. & Diananda, P. H. 1950. Extension of Quenouille's Test for Autoregressive Schemas. *Journal of the Royal Statistical Society, B* 12, 108ff.
- Bartlett, M. S. 1951. The Frequency Goodness of Fit Test for Probability Chains. *Proceedings of the Cambridge Philosophical Society*, 86-95.
- Bartlett, M. S. & Rajalkashman, D. V. 1953. Goodness-of-Fit Tests for Simultaneous Autoregressive Series. *Journal of the Royal Statistical Society*, 315, 107ff.
- Becker, P. & Schmidtke, A. 1977. Intelligenz und Hirnschädigung in ihrer Beziehung zur intellektuellen Lernfähigkeit. *Heilpädagogische Forschung*, VII.
- Bellman, R. & Astrom, K. J. 1970. On Structural Identifiability. *Mathematical Biosciences*, 7, 329-339.
- Bereiter, C. 1963. Some Persisting Dilemmas in the Measurement of Change. In: Harris, Ch. (Ed.): *Problems in Measuring Change*. Madison, Wisconsin.
- Bergstrom, A. R. & Wymer, C. R. 1976. A Model of Disequilibrium Neoclassical Growth and its Application to the UK. In: Bergstrom, A. R.: *Statistical Inference in Continous Time Models*. Amsterdam: North Holland Publ. Company, 267-327.
- Blalock, H. M. 1969. *Theory Construction: From Verbal to Mathematical Formulations*. Englewood Cliffs, N. J.: Prentice Hall.
- Bloomfield, P. 1976. *Fourier Analysis of Time Series: An Introduction*. New York: Wiley.
- Blumen, I., Kogan, M. & McCarthy, P. J. 1955. *The Industrial Mobility of Labor as a*

- Probability Process. Ithaka, New York: Cornell Studies of Industrial and Labor Relations, Volume 6.
- Bock, R. D. 1963. Multivariate Analysis of Variance of Repeated Measurements. In: Harris, Ch. (Ed.): Problems in Measuring Change. Madison, Wisc.
- Bock, R. D. 1975. Multivariate Statistical Methods in Behavioral Research. New York: Mc Graw-Hill.
- Bock, R. D. 1979. Univariate and Multivariate Analysis of Variance of Time-Structured Data. In: Nesselroade, J. R. & Baltes, P. B. (Eds.): Longitudinal Research in the Study of Behavior and Development. New York: Academic Press.
- Bortz, J. 1977. Lehrbuch der Statistik. Berlin: Springer.
- Bower, C. P., Padia, W. L. & Glass, G. V. 1974. TMS: Two Fortran IV Programms for the Analysis of Time-series Experiments. Boulder, Co.: Laboratory of Educational Research.
- Bower, G. H. & Trabasso, T. R. 1964. Concept Identification. Studies in Mathematical Psychology (R. C. Atkinson (Ed.)), Stanford.
- Box, G. E. P. 1949. A General Distribution Theory for a Class of Likelihood Criteria. *Biometrika*, 36, 317-346.
- Box, G. E. P. 1954. Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems: I. „Effect of Inequality of Variances in the One-Way Classification“. *Annals of Mathematical Statistics*, 25, 290-302.
- Box, G. E. P. & Jenkins, G. M. 1976<sup>3</sup>. Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.
- Box, G. E. P. & Pierce, D. A. 1970. Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time Series Models. *Journal of the American Statistical Association*, 64, 1509.
- Box, G. E. P. & Tiao, G. C. 1965. A Change in Level of a Non-Stationary Time Series. *Biometrika*, 52, 181-192.
- Box, G. E. P. & Tiao, G. C. 1973. Bayesian Inference in Statistical Analysis. Reading, Mass.: Addison-Wesley Publ. Co.
- Box, G. E. P. & Tiao, G. C. 1975. Intervention Analysis with Applications of Economic and Environmental Problems, *Journal of the American Statistical Association*, 70, 70-79.
- Box, G. E. P. & Tiao, G. C. 1977. A Canonical Analysis of Multiple Time Series. *Biometrika*, 64, 355ff.
- Bracht, G. H. & Glass, G. V. 1968. The External Validity of Experiments. *American Educational Research Journal*, 5, 437-474.
- Cadzow, J. A. & Martens, H. R. 1970. Discrete-Time and Computer Control Systems. Englewood Cliffs, N. J.: Prentice-Hall, Inc.
- Campbell, D. T. 1963. From Description to Experimentation: Interpreting Trends as Quasi-Experiments. In: Harris, Ch. (Ed.): Problems in Measuring Change. Madison, Wisc.

- Campbell, D. T. & Stanley, J. C. 1963. Experimental Designs for Research on Teaching. In: Gage, N. L., (Ed.): Handbook on Teaching. Chicago: Rand Mc Nally.
- Campbell, D. T. & Stanley, J. C. 1966. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand Mc Nally.
- Cattell, R. B. 1963. The Structuring of Change by P-Technique and Incremental R-Technique. In: Harris, Ch. (Ed.): Problems in Measuring Change. Madison, Wisc.
- Chan, S. P., Chan, S. Y. & Chan, S. G. 1972. Analysis of Linear Networks And Systems: A Matrix-Oriented Approach with Computer Applications. Reading, Mass.: Addison-Wesley.
- Chassan, J. B. 1979<sup>2</sup>. Research Design in Clinical Psychology and Psychiatry. New York: Halsted Press.
- Chatfield, C. 1975. The Analysis of Time Series: Theory and Practice. London: Chapman & Hall.
- Chow, G. C. 1975. Analysis and Control of Dynamit Economic Systems. New York: Wiley.
- Clauss, G., Guthke, J. & Lehwald, G. (Hrsg.), 1978. Psychologie und Psychodiagnostik lernaktiven Verhaltens. Berlin.
- Coleman, J. S. 1964. Introduction to Mathematical Sociology. New York: The Free Press.
- Coleman, J. S. 1966. Causal Models for qualitative Attributes. In: H. O. A. Wold (Ed.): Model Building in the Human Sciences. Monaco.
- Coleman, J. S. 1966. Causal Models for qualitative Attributes. In: H. O. A. Wold Blalock, A. B. (Eds.): Methodology in Social Research. New York: Mc Graw-Hill.
- Coleman, J. S. 1980. Unveröffentlichtes Manuskript.
- Cook, T. D. & Campbell, D. T. 1979. Quasi-Experiments: Nonequivalent Control Groups Designs. In: Cook & Campbell (Eds.): Quasi-Experimentation: Design & Analysis Issues for Field Settings, 95-146. Chicago: Rand Mc Nally.
- Cooley, W. W. & Lohnes, P. R. 1971. Multivariate Data Analysis. New York: Wiley.
- Conlisk, J. 1976. Interactive Markov Chains. Journal of Math. Sociology (4), 157-185.
- Cox, D. R. & Miller, H. D. 1968. The Theory of Stochastic Processes. London: Methuen.
- Cronbach, L. J. & Furby, L. 1970. How should we measure "Change" - or should we? Psychological Bulletin (74), 68-80.
- Da Silva, J. G. C. 1975. The Analysis of Cross-sectional Time Series Data. Institut of Statistics Mimeograph Series No. 1011. Raleigh, N. C.: North Carolina State University (Ph. D. Dissertation).
- Deppe, W. 1977. Formale Modelle in der Psychologie. Stuttgart: Kohlhammer.

- Doreian, P. & Hummon, N. P. 1976. *Modelling Social Processes*. Amsterdam: Elsevier.
- Doreian, P. & Hummon, N. P. 1977. Estimates for Differential Equation Models of Social Phenomena. In: Heise, P. R. (Ed.): *Sociological Methodology*, 180-208. San Francisco: Jossey-Bass Publ.
- Doreian, P. & Hummon, N. P. 1979. Estimating Differential Equation Models on Time Series: Some Simulation Evidence. *Sociological Methods & Research*, 8, 3-33.
- Drösler, J. 1976. Welche Zeiträume lassen sich mit psychologischen Prognosen überbrücken? In: Tack, W. H. (Hrsg.): *Bericht über den 30. Kongreß der Deutschen Gesellschaft für Psychologie in Regensburg*, 2, 3-14.
- Ebenhöh, W. 1975. *Mathematik für Biologen und Mediziner*. Heidelberg: Quelle & Meyer.
- Edgington, E. S. 1967. Statistical Inference from  $N = 1$  experiments. *The Journal of Psychology*, 65, 195-199.
- Edgington, E. S. 1969a. *Statistical Inference: The Distribution-free Approach*. New York: Mc Graw Hill.
- Edgington, E. S. 1969b. Approximate Randomization Tests. *Journal of Psychology*, 72, 143-149.
- Edgington, E. S. 1971. Randomization Tests with Statistical Control over Concomitant Variables. *Journal of Psychology*, 79, 13-19.
- Edgington, E. S. 1973. Randomization Tests: Computer Time Requirements. *Journal of Psychology*, 85, 89-95.
- Edgington, E. S. 1975a. Randomization Tests for One-Subject Operant Experiments. *Journal of Psychology*, 90, 57-68.
- Edgington, E. S. 1975b. Randomization for Predicted Trends. *Canadian Psychological Review*, 16, 49-53.
- Edgington, E. S. 1980. *Randomization Tests*. New York: Marcel Dekker.
- Elashoff, J. D. & Thoresen, C. E. 1978. Choosing a Statistical Method for Analysis of an Intensive Experiment. In: Kratochwill, Th. R. (Ed.), *Single Subject Research: Strategies for Evaluating Change*. New York: Academic Press, 287-311.
- Estes, W. K. & Suppes, P. 1974. Foundations of Stimulus Sampling Theory, In: Krantz, D. H., Atkinson, R. C., Luce, R. D., Suppes, P. (Eds.), *Contemporary Developments in Mathematical Psychology (Vol. 1): Learning, Memory and Thinking*, 163-184.
- Fararo, Th. J. 1973. *Mathematical Sociology*. New York: Wiley.
- Ferschl, F. 1970. *Markoffketten*. Berlin: Springer.
- Fichter, M. M. 1979. Versuchsplanung experimenteller Einzelfalluntersuchungen in der Psychotherapieforschung. In: Petermann, F. & Hehl (Hrsg.), *Einzelfallanalyse*. München: Urban & Schwarzenberg, 140-158.



- Finn, J. D. 1969. Multivariate Analysis of Repeated Measures Data, *Multivariate Behavioral Research*, 4, 391-413.
- Finn, J. D. 1974. *A General Model for Multivariate Analysis*. New York: Holt, Rinehart & Winston.
- Fischer, G. 1974. *Einführung in die Theorie psychologischer Tests*. Bern: Verlag Hans Huber.
- Fischer, G. H. 1976. Some Probabilistic Models for Measuring Change. In: De Gruiter, D. N. M. & van der Kamp, L. J. Th. (Eds.), *Advances in Psychological and Educational Measurement*. New York, 97-110.
- Fischer, G. H. 1977. Some Probabilistic Models for the Description of Attitudinal and Behavioral Changes under the Influence of Mass Communication. In: Kempf, W. F. & Repp, B. H. (Eds.), *Mathematical Models for Social Psychology*. Bern: Huber 102-151.
- Fischer, G. H. 1978. Probabilistic Test Models and their Applications. *The German Journal of Psychology*, 2, 298-319.
- Fisher, R. A. 1951<sup>6</sup>. *The Design of Experiments*, London: Hafner.
- Fisz, M. 1966, 1973<sup>2</sup>. *Wahrscheinlichkeitsrechnung und mathematische Statistik*. Berlin.
- Frame, J. S. 1964. Matrix Functions and Applications. *IEEE Spectrum*, Vol. 1, No. 6, June.
- Fuller, W. A. 1976. *Introduction to Statistical Time Series*. New York: Wiley.
- Fuller, W. A. & Battese, 1974. Estimation of Linear Models with Crossed Error Structure, *Journal of Econometrics*, 2, 67-78.
- Gaito, J. & Wiley, D. E. 1963. Univariate Analysis of Variance Procedures in the Measurement of Change. In: Harris, Ch. (Ed.), *Problems in Measuring Change*. Madison Wisconsin.
- Gastwirth, J. L. & Rubin, H. 1971. Effects of Dependence on the Level of Some One-Sample Tests. *Journal of the American Statistical Association*, 66, 816-820.
- Geisser, S. & Greenhouse, S. W. 1958. "An Extension of Box's Results on the Use of the F-Distribution in Multivariate Analysis". *The Annals of Mathematical Statistics*, 29, 885-891.
- Gentile, J. R., Roden, A. H. & Klein, R. D. 1972. An Analysis of Variance Model for the Intrasubject Replication Design. *Journal of Applied Behavior Analysis*, 5, 193-198.
- Ginsberg, R. B. 1971 (1). Semi-Markov Processes and Mobility. *Journal of Math. Sociology*, 233-263.
- Ginsberg, R. B. 1972 (2). Incorporating Causal Structure and Exogenous Information with Probabilistic Models: with Special Reference to Choice, Gravity, Migration and Markov Chains, *Journal of Math. Sociology*, 83-103.
- Glass, G. V., Peckham, P. D. & Sanders, J. R. 1972. Consequences of Failure to Meet Assumptions Underlying the Fixed-Effects Analyses of Variance and Covariance. *Review of Educational Research*, 42, 237-288.

- Glass, G. V., Willson, U. L. & Gottman, J. M. 1975. Design and Analysis of Time-Series Experiments. Boulder, Co.: Colorado Associated University Press.
- Goldberger, A. S. 1964. Econometric Theory. New York: Wiley.
- Goldstein, H. 1979. Some Models for Analysing Longitudinal Data on Educational Attainment. Journal of the Royal Statistical Soc., Series A, 142, 407-442.
- Goldstein, H. 1979. The Design and Analysis of Longitudinal Studies. New York: Academic Press.
- Goodman, L. A. 1961. Statistical Methods for the Mover-Stayer Model. Journal of the American Statistical Association, 56, 841-868.
- Goodman, L. A. 1974. Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models, Biometrika, 215-231.
- Gottman, J. M. 1981. Time Series Analysis. Cambridge: Cambridge University Press.
- Gottman, J. M. & Glass, G. V. 1978. Analysis of Interrupted Time-Series Experiments. In: Kratochwill, T. R. (Ed.), Single Subject Research: Strategies For Evaluating Change. New York: Academic Press, 197-235.
- Gottman, J. M. & Leiblum, S. R. 1974. How to do Psychotherapy and How to Evaluate it. New York: Holt, Rinehart & Winston.
- Granger, C. W. J. & Newbold, P. 1976. Identification of Two-way Causal Systems. In: Intriligator, M. D. (Ed.), Frontiers of Quantitative Economics II. Amsterdam.
- Granger, C. W. J. & Newbold, P. 1977. Forecasting Economic Time Series. New York.
- Greenhouse, S. W. & Geisser, S. 1959. On Methods in the Analysis of Profile Data. Psychometrika, 24, 95-112.
- Greeno, J. G. 1974. Representation of Learning as Discrete Transition in a Finite State Space. In: Krantz, D. H. et al. (Eds), Contemporary Developments in Mathematical Psychology: Learning Memory and Thinking, 1-43, San Francisco: Freeman.
- Grizzle, J. & Allen, D. M. 1969. Analysis of Growth and Dose Response Curves. Biometrics, 25, 357-381.
- Gröbner, W. 1966. Matrizenrechnung. Mannheim: BI-Verlag.
- Guthke, J. 1976. Entwicklungsstand und Probleme der Lernfähigkeitsdiagnostik, Teil I/II. Zeitschrift für Psychologie.
- Guthke, J. 1977<sup>3</sup>. Zur Diagnostik der intellektuellen Lernfähigkeit. Berlin.
- Guthke, J. 1980. Die Relevanz des Lerntestkonzepts für die klinisch-psychologische Diagnostik - demonstriert am Beispiel der Diagnostik der geistigen Behinderung und der frühkindlichen Hirnschädigung. Probleme und Ergebnisse der Psychologie, 72, 5-21.
- Hall, B. H. 1978. A General Framework for the Time Series-Cross Section Estimation. Annales de L'Insee Nr. 30-31, 177-202.
- Hall, R. V., Fox, R., Willard, D., Goldsmith, L., Emerson, M., Owen, M., Davis, F.

- & Porcia, E. 1971. The Teacher as Observer and Experimenter in the Modification of Disputing and Talking-out Behaviors. *Journal of Applied Behavior Analysis*, 4, 141-149.
- Hamouzova, M. & Würthner, K. 1976. Der Einfluß zweier Trainingsmethoden auf die Intelligenztestleistung Erwachsener. Unveröfftl. Diplomarbeit, Psychologisches Institut der Universität Heidelberg.
- Hannan, E. J. 1970. *Multiple Time Series*. New York: J. Wiley & Sons.
- Hannan, E. J. 1976. The Identification and Parametrization of ARMAX and State Space Forms. *Economics*, 44, 713-722.
- Hannan, M. T. & Young, A. A. 1977. Estimation in Panel Models: Results on Pooling Cross-Sections and Time Series. *Sociological Methodology*, 52-82.
- Harder, Th. 1973. *Dynamische Modelle in der empirischen Sozialforschung*. Stuttgart, Teubner.
- Harris, Ch. W. (Ed.), 1963. *Problems in Measuring Change*. The University of Wisconsin Press.
- Harris, Ch. 1963. Canonical Factor Models for the Description of Change. In: Harris, Ch. (Ed.), *Problems in Measuring Change*. Madison Wisconsin.
- Hart, M. C. & Mulholland, R. J. 1979. Structural Identifiability of Compartmental Systems Based upon Measurements of Accumulated Tracer in Closed Pools. *Mathematical Biosciences*, 47, 239-253.
- Hartmann, D. P. 1974. Forcing Square Pegs Into Round Holes: Some Comments on "An Analysis-of-Variance Model for the Intrasubject Replication Design". *Journal of Applied Behavior Analysis*, 7, 635-638.
- Helmer, R. M. & Johansson, J. K. 1977. An Exposition of the Box-Jenkins Transfer Function Analysis with an Application to the Advertising-Sales Relationships. *Journal of Marketing Research*, 14, 227-239.
- Henderson, C. R. 1971. Comment on "The Use of Error Component Models in Combining Cross Section with Time Series Data". *Econometrica* (39), 397-401.
- Henry, N. W. 1973. Measurement Models for Continuous and Discrete Variables. In: Goldberger, A. S. & Duncan O. D. (Eds), *Structural Equation Models in the Social Sciences*. New York: Seminar Press.
- Hersen, M. & Barlow, D. H. 1976. *Single-Case Experimental Designs: Strategies for Studying Behavior Change*. New York: Pergamon Press.
- Hibbs, D. A. 1974. Problems of Statistical Estimation and Causal Inference in Time-Series Regression Models. In: Costner, H. S. (Ed.), *Sociological Methodology*, 1973-1974, 252-308. San Francisco: Jossey-Bass Publishers.
- Hibbs, D. A. Jr. 1977. On Analyzing the Effects of Policy Interventions: BOX-JENKINS and BOX-TIAO vs. Structural Equation Models. In: Heise, D. R. (Ed.), *Sociological Methodology*, 137-179. New York: Jossey Bass.
- Hilgard, J. R. 1933. The Effect of Delayed Practice on Memory and Motor Performance Studied by the Method of Co-twin Control. *Genetic Psychology Monographs*, 6, 67ff.

- Holtzman, W. H. 1963. Statistical Models for the Study of Change in the Single Case. In: Harris, Ch. (Ed.), Problems in Measuring Change. Madison Wisconsin.
- Horan, P. H. 1976. Structure and Change in Occupational Mobility: a Markov Approach. *Quality & Quantity*, 10, 321-340.
- Horst, P. 1963. Multivariate Models for Evaluating Change. In: Harris, Ch. (Ed.), Problems in Measuring Change. Madison Wisconsin.
- Howard, R. A. 1971. Dynamic Probabilistic Systems, Vol. 1: Markov Models, Vol. II: Semimarkov and Decision Processes. New York: Wiley.
- Huba, G. J., Lawler, W. G., Stallone, F. & Fieve, R. R. 1976. The Use of Autocorrelation Analysis in the Longitudinal Study of Mood Patterns in Depressed Patients. *British Journal of Psychiatry*, 128, 146-155.
- Hummon, N. P., Doreian, P. & Teuter, K. 1975. A Structural Control Model of Organizational Change. *American Sociological Review*, 40, 813-824.
- Huynh, Huynh & Feldt, L.S. 1970. Conditions under which Mean Square Ratios in Repeated Measurements Designs Have Exact F-Distributions. *Journal of the American Statistical Association*, 65, 1582-1589.
- Isaak, P. D. 1970. Linear Regression, Structural Relations and Measurement Error. *Psychological Bulletin*, 73, 213-218.
- Jenkins, G. M. 1979. Practical Experiences with Modeling and Forecasting Time Series. Jersey, Channel Island: GJP Ltd.
- Jöreskog, K. G. 1973. A General Method for Estimating a Linear Structural Equation System. In: Goldberger, A. S. & Duncan, O. D. (Eds), *Structural Equation Models in The Social Sciences*. New York: Seminar Press, 85-112.
- Jöreskog, K. G. 1979. Statistical Estimation of Structural Models in Longitudinal Developmental Investigations. In: Nesselroade, J. R. & Baltes, P. B. (Eds), *Longitudinal Research in the Study of Behavior and Development*. New York: Academic Press, 303-351.
- Jöreskog, K. G. & Sörbom, D. 1976. Statistical Models and Methods for Test-Retest-Situations. In: De Gruijter, D. N. M. & van der Kamp, L.J. Th. (Eds), *Advances in Psychological and Educational Measurement*. New York, 135-157.
- Jöreskog, K. G. & Sörbom, D. 1977. Statistical Models and Methods for Analysis of Longitudinal Data. In: Aigner, P. J. & Goldberger, A. S. (Eds), *Latent Variables in Socio-Economic-Models*. Amsterdam, 285-325.
- Kazdin, A. E. 1976. Statistical Analysis for Single-Case Experimental Designs. In: Hersen, M. & Barlow, D. H. (Eds), *Single-Case Experimental Designs: Strategies for Studying Behavior Change*. New York: Pergamon Press.
- Kaiser, H. F. 1963. Image Analysis. In: Harris, Ch. (Ed.), Problems in Measuring Change. Madison Wisconsin.
- Kemeny, J. G. & Snell, I. L. 1965. *Finite Markov Chains*. New York.
- Kendall, M. G. & Stuart, A. 1973. *The Advanced Theory of Statistics II*. London: Griffin & Co., 3. Auflage.

- Keselman, H. J. & Leventhal, L. 1974. Concerning the Statistical Procedures Enumerated by Gentile et al.: Another Perspective. *Journal of Applied Behavior Analysis*, 7, 632-645.
- Khatri, C. G. 1966. A Note on a MANOVA Model Applied to Problems in Growth Curve. *Annals of the Institute of Statistical Mathematics*, 18, 75-86.
- Kleiter, E. 1979. Clusteranalytische Struktur-Veränderungsmessung zur Abbildung von Lernen als Struktur-Veränderung. Vortrag auf dem IPN-Seminar „Veränderungsmessung zur Diagnose und Prognose von Lerneffekten“, 1.-5. Oktober 1979 im Institut für Pädagogik der Naturwissenschaften, Kiel.
- Kleiter, E. & Petermann, F. 1977. *Abbildung von Lernwegen*. München: Oldenbourg.
- Kormann, A. 1981. Veränderungsmessung. In: Schiefele, H. & Krapp, A. (Eds), *Handlexikon der Pädagogischen Psychologie*. München: Ehrenwirth.
- Kormann, A. 1979. Lerntests - Versuch einer kritischen Bestandsaufnahme. In: Eckensberger, L. H. (Ed.), *Bericht über den 31. Kongreß der Deutschen Gesellschaft für Psychologie*, Göttingen, Bd. 2, 85-95.
- Krapp, A. & Schiefele, H. 1976. *Lebensalter und Intelligenzentwicklung*. München.
- Kuhl, J. & Blankenship, V. 1979. The Dynamic Theory of Achievement Motivation: From Episodic To Dynamic Thinking. *Psychological Review*, 86, 141-151.
- Kuriakjian, B. & Zelen, M. 1962. A Calculus for Factorial Arrangements. *Annales of Math. Statistics* (33), 609-619.
- Laming, D. 1973. *Mathematical Psychology*. London: Academic Press.
- Land, K. C. 1970. Mathematical Formalization of Durkheim's Theory of Division of Labor. In: Borrgatta, E. F. & Bohrnstedt, G. W. (Eds), *Sociological Methodology 1970*. San Francisco: Jossey-Bass Inc., 257-282.
- Land, K. C. 1971. Formal Theory. In: Costner, H. L. (Ed.), *Sociological Methodology 1971*, 175-220. San Francisco: Jossey-Bass Inc.
- Lazarsfeld, O. & Henry, N. W. 1968. *Latent Structure Analysis*. Boston (Mass.): Houghton Mifflin.
- Lee, T. C., Judge, G. G. & Zehner, A. 1977. *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data*. Amsterdam: North-Holland.
- Lehman, E. L. 1975. *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.
- Levin, J. R., Marascuilo, L. A. & Hubert, L. J. 1978. N = Nonparametric Randomization Tests. In: Kratochwill, Th. R. (Ed.), *Single Subject Research*. New York: Academic Press, 167-196.
- Levine, G. & Burke, C. J. 1972. *Mathematical Model Techniques for Learning Theories*. New York: Academic Press.
- Lewandowski, R. 1980. *Prognose- und Informationssysteme*. Berlin: W. de Gruyter.
- Löhr, H. J. 1979. *Beispiele und Aufgaben zur Laplace-Transformation*. Braunschweig: Vieweg & Sohn.

- Long, J. S. 1976. Estimation and Hypothesis Testing in Linear Models Containing Measurement Error: A Review of Jöreskog's Model for the Analysis of Covariance Structures. *Sociological Methods & Research*, 5, 157-206.
- Loose, K. D. 1964. Using an Energy Systems Approach in Modeling Achievement Motivation and Some Related Sociological Concepts. Unpublished Doctoral Dissertation. The University of Florida, August 1964.
- Loose, K. D. & Koran, J. J. Jr. 1975. A Procedure for the Investigation of Hypothesized Relationships Among Dynamic Variables Over Time. Paper presented at the Annual Meeting of the American Educational Research Association. Washington, D. C., April 1975.
- Loose, K. D. & Unruh, W. R. 1977. An Analysis of Interaction of Anxiety, Aspiration Level And Ability Derived From Ecological Systems Theory. Paper presented at the Annual Meeting of the American Educational Research Association. San Francisco, April 1976 und Reviews In Education.
- Loose, K. D. & Unruh, W. R. 1977. An Ecological System Model For Increasing Predictive Validity by Using Variables Frequently Excluded from Regression Analysis. A Paper Presented at the Third International Symposium on Educational Testing, Leyden, June 1977.
- Lord, F. M. 1963. Elementary Models for Measuring Change. In: Harris, Ch. (Ed.), *Problems in Measuring Change*. Madison Wisconsin.
- Makridakis, S. & Wheelwright, S. C. 1978a. *Interactive Forecasting: Univariate and Multivariate Methods*. San Francisco: Holden-Day.
- Makridakis, S. & Wheelwright, S. C. 1978b. *Forecasting: Methods and Applications*. New York: John Wiley.
- Malinvaud, E. 1970 (2). *Statistical Methods of Econometrics*. Amsterdam: North Holland.
- Markus, G. B. & Zajonc, R. B. 1977. Systems Simulation of Family Configuration and Intellectual Development: A Simulation. *Behavioral Science*, 22, 137-142.
- Marmor, Y. S. & Marmor, M. 1978. Comment on Simonton's Cross-Sectional Time Series Experiments: Some Suggested Statistical Analysis. *Psychological Bulletin*, 85, 1102-1103.
- Massy, W. F., Montgomery, D. B. & Morrison, D. Y. 1970. *Stochastic Models of Buying Behavior*. Cambridge, Mass.: The M.I.T. Press.
- McCain, L. J. & McCleary, R. 1979. The Statistical Analysis of the Simple Interrupted Time-Series Quasi-Experiment. In: Cook, Th. D. & Campbell, D. T. (Eds), *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago: Rand McNally College Publishing Co., 233-293.
- McCall, R. B. & Applebaum, M. I. 1973. Bias in the Analysis of Repeated Measures Designs: Some Alternative Approaches, *Child Development*, 44, 401-415.
- McCleary, R. & Hay, R. A. 1980. *Applied Time Series Analysis For The Social Sciences*, Beverly Hills: Sage Publications.

- McClelland, J. L. 1979. On the Time Relations of Mental Processes: An Examination of Systems of Processes in Cascade. *Psychological Review*, 86, 287-330.
- McDonald, R. P. & Swaminathan, H. 1973. A Simple Matrix Calculus with Applications to Multivariate Analysis. *General Systems*, 18, 37-54.
- McGill, W. J. Stochastic Latency Mechanisms. In: Luce, R. D., Bush, R. R. & Galanter, E. (Eds), *Handbook of Mathematical Psychology*, Vol. 1, 309-360.
- McGinnis, R. 1968. A Stochastic Model of Social Mobility. *American Sociological Review*, 33, 712-722.
- McNeil, K. A., Kelly, F. J. & McNeil, J. T. 1975. Testing Research Hypothesis Using Multiple Linear Regression. Carbondale and Edwardsville: Southern Illinois University Press.
- McSweeney, A. J. 1978. The Effects of Response Cost on the Behavior of a Million Persons: Charging for Directory Assistance in Cincinnati. *Journal of Applied Behavioral Analysis*, 11, 47-51.
- Meier, F. 1981. Zur Gewinnung und Bedeutungszuordnung personspezifischer Prozeßparameter der Befindlichkeit. *Diagnostica*, 27, 23-38.
- Melchinger, H. 1978. Intelligenz als Lernfähigkeit: Ein empirischer Beitrag zum Lern-testkonzept. Inaug. Dissertation am Institut für Psychologie (FB 12) der FU Berlin.
- Metz-Göckel, H. 1979. Das Messen von Veränderungen in Einstellung und Verhalten. In: Heinerth, K. (Ed.), *Einstellung und Verhalten*, München, 356-389.
- Miller, A. D. 1971. Logic of Causal Analysis: From Experimental to Nonexperimental Designs. In: Blalock, H. M. (Ed.), *Causal Models in the Social Sciences*. Chicago: Aldine, 273-294.
- Miller, G. A. & Chomsky, N. 1963. Finitary Models of Language Users. In: Luce, R. D., Bush, R. R. & Galanter, E. (Eds), *Handbook of Mathematical Psychology*. Vol. 11, 419-491.
- Möbus, C. 1981. Zur Beschreibung und Analyse kurz- und langfristiger Testintelligenzveränderungen mit zeitdiskreten und zeitkontinuierlichen dynamischen Modellen. In: Michaelis (Ed.), *Bericht über den 32. Kongreß der Deutschen Gesellschaft für Psychologie*, Zürich: 22. 09. 80-25. 09. 80. Göttingen.
- Möbus, C. 1983. Identification and Fitting of Continuous Time Models with Two-Wave Data (in preparation).
- Möbus, C., Göricke, G. & Kröh, P. 1982. Statistical Analysis of Single-Case Experimental Designs: Conditional Equivalence of the General-Linear-Model Approach of Glass, Willson & Gottman with the Intervention Model of Box & Tiao (in press).
- Möbus, C. & Wallasch, R. 1977. Zur Erfassung von Hirnschädigungen bei Kindern: Nichtlineare Entscheidungsregeln auf der Basis von Veränderungsmessungen und des Jackknife. *Diagnostica*, 227-251.
- Moosbrugger, H. 1978. *Multivariate Statistische Analyseverfahren*. Stuttgart: Kohlhammer.

- Morrison, D. F. 1976<sup>2</sup>. Multivariate Statistical Methods. New York: McGraw Hill.
- Mundlak, Y. 1978. On The Pooling of Time Series and Cross Section Data. *Econometrica*, 46, 69-85.
- Murray, J. R., Wiley, D. E. & Wolfe, R. G. 1971. New Statistical Techniques for Evaluating Longitudinal Models. *Human Development*, 14, 142-148.
- Nagl, W. H. 1983. Längsschnittmethoden. Arbeitsbericht des Zentrums I, Universität Konstanz.
- Namboodiri, N. K., Carter, L. F. & Blalock, H. M. 1975. Applied Multivariate Analysis And Experimental Designs. New York: McGraw Hill.
- Nelson, C. R. 1973. Applied Time Series Analysis. San Francisco: Holden Day, Inc.
- Nerlove, M. 1971a. A Note on Error Components Models. *Econometrica*, 39, 383-396.
- Nerlove, M. 1971b. Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections. *Econometrica*, 39, 359-382.
- Nelson, Ch. R. 1973. Applied Time Series Analysis. San Francisco: Holden-Day.
- Nesselroade, J. R. & Baltes, P. B. 1979. Longitudinal Research in the Study of Behavior and Development. New York: Academic Press.
- Neymann, J. & Scott, E. L. 1948. Consistent Estimates Based on Partially Consistent Observations. *Econometrica*, 16, 1-10.
- Pack, D.J. 1978. A Computer Program for the Analysis of Time-Series-Models Using the Box-Jenkins Philosophy. Hatboro, Pa.: Automatic Forecasting Systems.
- Parks, R. W. 1967. Efficient Estimation of a System of Regression Equations when Disturbances are Both Serially and Contemporaneously Correlated. *Journal of the American Association*, 62, 500-509.
- Pawlik, K. (Ed.). 1976. Diagnose der Diagnostik. Stuttgart: Klett.
- Pedhazur, E. J. 1977. Coding Subjects in Repeated Measures Designs. *Psychological Bulletin*, 84, 298-305.
- Petermann, F. 1978. Veränderungsmessung. Stuttgart: Kohlhammer.
- Petermann, F. & Hehl, F. J. (Eds), 1979. Einzelfallanalyse. München.
- Pfeiffer, P. E. & Deutsch, S. J. 1980. A Three Stage Iterative Procedure for Space-Time Modeling. *Technometrics*, 22, 35-47.
- Phillips, P. C. B. 1972. The Structural Estimation of a Stochastic Differential Equation System. *Econometrica*, 40, 1021-1041.
- Phillips, P. C. B. & Wickens, M. R. 1978. Exercises In Econometrics, Vol. II. Oxford: Philip Allan Publishers Ltd.
- Potthoff, R. F. & Roy, S. N. 1964. A Generalized Multivariate Analysis of Variance Model Useful Especially in Growth Curve Problems. *Biometrika*, 51, 313-326.
- Priestley, M. D. 1980. System Identification, Kalman Filtering and Stochastic Control. Paper presented at the Conference of the Institute of Mathematical Statistics on Time Series Analysis. JA, 1978. Zitiert nach SAS User's Guide.



- Przeworski, A. & Soares, G. A. D. 1977. Theorien auf der Suche nach einer Kurve: Strukturelle Interpretation der Linkswahl. In: Schmutzer, M. E. A. (Ed.), *Mathematische Methoden in der Politikwissenschaft*. München: Oldenbourg, 31-70.
- Quenouille, M. H. 1949. Approximate Tests of Correlation in Time Series. *Journal of the Royal Statistical Society*, B 11, 68.
- Quenouille, M. H. 1957. *The Analysis of Multiple Time Series*. London.
- Ralston, M. L., Jennrich, R. J., Sampson, P. F. & Kuo, F. K. 1979. Fitting Pharmacokinetic Models with BMDPAR, BMDP Technical Report No. 58.
- Rao, C. R. 1959. Some Problems Involving Linear Hypothesis in Multivariate Analysis. *Biometrika*, 46, 49-58.
- Rao, C. R. 1965. The Theory of Least Squares When the Parameters are Stochastic and its Application to the Analysis of Growth Curves. *Biometrika*, 52, 447-458.
- Rao, C. R. 1966. Covariance Adjustment und Related Problems in Multivariate Analysis. In: Krishnaiah, P. R. (Ed.), *Multivariate Analysis I*. New York: Academic Press.
- Rao, C. R. 1967. Least Square Using an Estimated Dispersion Matrix and its Application to Measurements of Signals. *Proceedings of the 5th Berkely Symposium on mathematical statistics*. Berkely: University of California Press.
- Rapoport, A. 1963. Mathematical Models of Social Interaction. In: Luce, R. D., Bush, R. R. & Galanter, E. (Eds), *Handbook of Math. Psychology*, Vol. II, 493-578.
- Rapoport, A. 1980. *Mathematische Methoden in den Sozialwissenschaften*. Würzburg: Physica-Verlag.
- Rashevsky, N. 1939. Studies in the Mathematical Theory of Human Relations. *Psychometrika*, 4, 221-239.
- Reichardt, Ch. S. 1979. The Statistical Analysis of Data from Nonequivalent Group Designs. In: Cook & Campbell (Eds), *Quasi-Experimentation. Design & Analysis Issues for Field Settings*, 147-205. Chicago: Rand McNally.
- Renn, H. 1973. *Die Messung von Sozialisierungswirkungen*. München.
- Restle, F. & Greeno, J. G. 1970. *Introduction to Mathematical Psychology*. Reading, Mass.: Addison-Wesley.
- Revenstorf, D. 1979. *Zeitreihenanalyse für klinische Daten: Methodik und Anwendungen*. Weinheim: Beltz Verlag.
- Revenstorf, D. & Keeser, W. 1979. *Zeitreihenanalyse von Therapieverläufen*. In: Petermann, F. & Hehl, F. J. (Eds), *Einzelfallanalyse*. München: Urban & Schwarzenberg, 183-228.
- Revenstorf, D. & Vogel, B. 1979. Zur Analyse qualitativer Verlaufsdaten - ein Überblick. In: Petermann, F. & Hehl, F. J. (Eds), *Einzelfallanalyse*, 229-250.
- Revenstorf, D., Wegschneider, R., Fitting, U. & Mai, N. 1977. Markov Models of Gaming Behavior in Experimental Non-Zero-Sum Games. In: Kempf, W. F. & Repo, B. H. (Eds), *Mathematical Models for Social Psychology*. Bern: Huber.

- Richardson, L. F. 1948. War Moods. *Psychometrika*, 13, 147-174.
- Ronning, G. 1980. ökonometrische Analyse von aggregierten Tendenzdaten aus Panelerhebungen. Vortrag im statistischen Seminar. Universität Konstanz.
- Roskam, E. E. 1976. Multivariate Analysis of Change and Growth: Critical Review and Perspectives. In: Gruijter, D. N. M. & van der Kamp, L. J. Th. (Eds), *Advances in Psychological and Educational Measurement*. New York: J. Wiley, 111-133.
- Roskam, E. E. 1979. Varianzanalytische Verfahren und Strukturmodelle für Längsschnittdaten. Vortrag auf dem Seminar „Veränderungsmessung zur Diagnose und Prognose von Lerneffekten“ am Institut der Pädagogik der Naturwissenschaften, 01. 05.-05. 10. 1979.
- Rost, J. & Spada, H. 1978. Probabilistische Testtheorie. In: Klauer, K. J. (Ed.), *Handbuch der Pädagogischen Diagnostik*, Bd. 1. Düsseldorf, 59-97.
- Rudinger, G. & Lantermann, E. D. 1978. Probleme der Veränderungsmessung in individuellen und gruppentypischen Entwicklungsverläufen. In: Oerter, R. (Ed.), *Entwicklung als lebenslanger Prozeß*. Frankfurt/M., 178-226.
- Scheffé, H. 1959. *The Analysis of Variance*. New York: John Wiley & Sons.
- Schuss, Z. 1980. *Theory and Applications of Stochastic Differential Equations*, New York: Wiley.
- Schweitzer, W. 1978. Modelle zur Erfassung von Wanderungsbewegungen. Meisenheim am Glan: Verlag Anton Hain.
- Searle, S. R. 1971. *Linear Models*. New York: Wiley.
- Shampine, L. F., Watts, H. A. & Davenport, S. M. 1976. Solving Nonstiff Ordinary Differential Equations - the State of the Art. *SIAM Review*, 18, 376-441.
- Shine, L. C. & Bower, S. M. 1971. A One-Way Analysis of Variance for Single-Subject Designs. *Educational and Psychological Measurement*, 31, 105-113.
- Simon, H. A. 1952. A Formal Theory of Interaction in Social Groups. *Am. Sociological Review*, 17, 202-211.
- Simon, H. A. 1957. *Models of Man*. New York: Wiley.
- Simon, H. A. & Newell, A. 1974. Thinking Processes. In: Krantz, D. H., Atkinson, R. C., Luce, R. D. & Suppes, P. (Eds): *Contemporary Developments in Mathematical Psychology. Learning, Memory and Thinking*. San Francisco.
- Simonton, D. K. 1977. Cross-Sectional Time Series Experiments: Some Suggested Statistical Analyses. *Psychological Bulletin*, 84, 489-502.
- Singer, B. 1981. Estimation of Nonstationary Markov Chains from Panel Data, In: Leinhardt, S. (Ed.): *Sociological Methodology*. San Francisco: Jossey-Bass, 319-337.
- Singer, B. & Spilerman, S. 1976a. Representation of Social Processes by Markov Models. *American Journal of Sociology*, 82, 1-53.
- Singer, B. & Spilerman, S. 1976b. Some Methodological Issues in the Analysis of Longitudinal Surveys. *Annales of Economic & Social Measurement*, 447ff.

- Singer, B. & Spilerman, S. 1979a. Mathematical Representations of Development Theories. In: Nesselroade, J. R. & Baltes, P. B. (Eds): *Longitudinal Research in the Study of Behavior and Development*. New York: Academic Press, 155-177.
- Singer, B. & Spilerman, S. 1979b. Clustering on the Main Diagonal in Mobility Matrices. In: Schuessler, K. F. (Ed.): *Sociological Methodology*. San Francisco, 172-208.
- Slutzky, E. 1937. The Summation of Random Causes as the Source of Cyclic Processes. *Econometrica*, 5, 105-146.
- Sörbom, D. 1976. A Statistical Model for the Measurement of Change in True Scores. In: De Gruijter, D. N. M. & van der Kamp, L. J. Th. (Eds): *Advances in Psychological and Educational Measurement*. New York, 159-170.
- Sörbom, D. 1979. LISREL IV with Structured Means. Workshop „Lineare Strukturgleichungsmodelle“ am Zentrum für Umfragen und Analysen, ZUMA in Mannheim, Oktober 1979.
- Spilerman, S. 1972. Extensions of the Mover-Stayer Model. *American Journal of Sociology*, 78, 599-626.
- Spilerman, S. 1972. The Analysis of Mobility Processes by the Introduction of Independent Variables into a Markov Chain. *American Sociological Review*, 37, 277-294.
- Steyer, R. 1980. Stochastische Prozesse und kausale Abhängigkeit. Vortrag auf dem XXII. Internationalen Kongreß für Psychologie, Leipzig, 6.-12. Juli 1980.
- Storm, R. 1969. *Wahrscheinlichkeitsrechnung*. Leipzig: VEB Fachbuchverlag.
- Straka, G. 1974. *Forschungsstrategien zur Evaluation von Schulversuchen*. Weinheim.
- Swaminathan, H. & Algina, J. 1977. Analysis of Quasi-experimental Time-series Designs. *Journal of Multivariate Behavioral Research*, 12, 111-131.
- Tack, W. H. 1976. *Stochastische Lernmodelle*. Stuttgart: Kohlhammer.
- Thalmaier, A. 1979. Zur kognitiven Bewältigung der optimalen Steuerung eines dynamischen Systems. *Zeitschrift für experimentelle und angewandte Psychologie*, 26, 388-421.
- Theil, H. 1971. *Principles of Econometrics*. New York: Wiley.
- Thoresen, C. E. & Elashoff, J. D. 1974. An Analysis of Variance Model for Intrasubject Replication Design: Some Additional Comments. *Journal of Applied Behavior Analysis*, 7, 639-642.
- Timm, N. H. 1975. *Multivariate Analysis with Applications in Education and Psychology*. Monterey, Calif.: Brooks/Cole Publ. Co.
- Tucker, L. R. 1963. Implications of Factor Analysis of Three-Way Matrices for Measurement of Change. In: Harris, Ch. (Ed.): *Problems in Measuring Change*. Madison, Wisconsin.
- Tuma, N. B., Hannan, M. T. & Groenveld, L. P. 1979. Dynamic Analysis of Event Histories. *American Journal of Sociology*, (84), 820-854.
- Tuma, N. B. & Hannan, M. T. 1979. Approaches to the Censoring Problems in

- Analysis of Event Histories. In: Schuessler, K. F. (Ed.): *Sociological Methodology*. San Francisco. 209-240.
- Tuma, N. B. 1980. When Can Interdependence In A Dynamic System of Qualitative Variables be Ignored? In: Schuessler, K. F. (Ed.): *Sociological Methodology*. San Francisco, 358-391.
- Tyler, V. D. & Brown, G. D. 1968. Token Reinforcement of Academic Performance with Institutionalized Delinquent Boys. *Journal of Educational Psychology*, 59, 164-168.
- Vigderhous, G. 1978. Forecasting Sociological Phenomena: Application of Box-Jenkins Methodology to Suicide Rates. In: Schuessler, K. F. (Ed.): *Sociological Methodology*, 20-51, San Francisco: Jossey-Bass.
- Voevodsky, J. 1969. Quantitative Behavior of Warring Nations. *Journal of Psychology*, 72, 269-292.
- Walker, A. M. 1931. On the Periodicity in Series of Related Terms. *Proceedings of the Royal Society of London*, A 131, 518-532.
- Wallace, T. D. & Hussain, A. 1969. The Use of Error Components Models in Combining Cross Section with Time Series Data. *Econometrica*, (37), 55-72.
- Wasserman, S. S. 1980. A Stochastic Model for Directed Graphs with Transition Rates Determined by Reciprocity. In: Schuessler, K. F. (Ed.): *Sociological Methodology*. San Francisco: Jossey-Bass, 392-412.
- Webster, H. & Bereiter, C. 1963. The Reliability of Changes Measured by Mental Test Scores. In: Harris, Ch. (Ed.): *Problems in Measuring Change*. Madison, Wisconsin.
- Werts, C. E. & Linn, R. L. 1970. A General Linear Model for Studying Growth. *Psychological Bulletin*, 73, 17-22.
- Werts, C. E. & Linn, R. L. 1971. Considerations When Making Inferences Within the Analysis of Covariance model. *Educational and Psychological Measurement*, 31, 407-416.
- Werts, C. E. & Linn, R. L. 1972. Corrections for Attenuation. *Educational and Psychological Measurement*, (32), 117-127.
- Werts, C. E. & Jöreskog, K. G. & Linn, R. L. 1972. A Multitrait-Multimethod Model for Studying Growth. *Educational and Psychological Measurement*, (32), 655-678.
- Wheaton, B., Mutherr, B., Alwin, D. F., Summers, G. F. 1977. Assessing Reliability and Stability in Panel Models. *Sociological Methodology*, 84-136.
- Wiggins, L. M. 1973. *Panel Analysis: Latent Probability Models for Attitude and Behavior Processes*. Amsterdam: Elsevier.
- Wiley, D. E. & Wiley, J. A. 1970. The Estimation of Measurement Error in Panel Data. *American Sociological Review*, (35), 112-117.
- Wilks, S. S. 1932. Certain Generalizations in the Analysis of Variance, *Biometrika*, 24, 471-494.

- Wilson, G. T. 1973. The Estimation of Parameters in Multivariate Time Series models. *Journal of the Royal Statistical Society, B* 35, 76ff.
- Winer, B. J. 1971. *Statistical Principles in Experimental Design*. New York: Mc Graw Hill.
- Wottawa, H. 1974. Das Allgemeine Lineare Modell - Ein universelles Auswertungsverfahren. *EDV in Medizin und Biologie*, 3, 65-73.
- Wottawa, H. 1979. *Grundlagen und Probleme von Dimensionen in der Psychologie*. Meisenheim: Hain.
- Yule, G. U. 1926. Why Do We Sometimes Get Nonsense-Correlations Between Time Series? A Study in Sampling and the Nature of Time Series. *Journal of the Royal Statistical Society*, 89, 1-64.
- Yule, G. U. & Kendall, M. G. 1964. *An Introduction to the Theory of Statistics*. London: Griffin (4. Auflage).
- Zelen, M. & Federer, W. T. 1964. Applications of the Calculus for Factorial Arrangements. II Designs with Twoway Elimination of Heterogeneity. *Annales of Math. Statistics*, 35, 658-672.
- Ziegler, R. 1972. *Theorie und Modell*. München: Oldenbourg.
- Zimmermann, P. 1979. Zur Zeitreihenanalyse von Stimmungsskalen. *Diagnostica*, 25, 24-48.
- Zurmühl, R. 1964<sup>4</sup>. *Matrizen*. Berlin: Springer-Verlag.

## 4. Kapitel

# Statistische Entscheidungstheorie und Bayes-Statistik

*Dirk Wendt*

### *1. Einleitung: Problemstellung*

Wissenschaften bestehen im allgemeinen aus einer Menge von Aussagen über ihren Gegenstand - im Falle der Psychologie handelt es sich um empirisch fundierte Aussagen über das Verhalten (und Erleben) von Menschen.

Eine wichtige Aufgabe jeder Wissenschaft ist es, Kriterien aufzustellen, denen Aussagen genügen müssen, um in den Bestand des „Wissens“, der die Wissenschaft aufmacht, aufgenommen werden zu können: Aus der Vielfalt möglicher Aussagen müssen die herausgefiltert werden, die diesen Anforderungen entsprechen und damit als „gesichertes Wissen“ oder in gewissem Sinne als „wahr“ (genauer genommen: bestätigt) gelten können.

Wir müssen also Entscheidungen darüber treffen, welche Sätze in den Kanon des „Wissens“ aufgenommen werden sollen und welche nicht. Aus Gründen, die in diesem Aufsatz näher erläutert werden, können solche Entscheidungen in einer empirischen Wissenschaft wie der Psychologie immer nur vorbehaltlich einer gewissen Irrtumswahrscheinlichkeit getroffen werden.

Wir wollen in diesem Kapitel einige Strategien besprechen, die sich in der Psychologie und ihren Nachbardisziplinen eingebürgert haben, die durch Konventionen akzeptiert sind und zu gewissen Erfolgen dieser Wissenschaften bei der Erklärung und Vorhersage der von ihnen behandelten Phänomene geführt haben. In gewissem Sinne handelt es sich bei jeder empirischen, d.h. auf Erfahrungen aufbauenden, Wissenschaft um eine Abbildung oder Repräsentation beobachtbarer Sachverhalte durch Symbole, meistens durch die Worte der Sprache. Ein beobachteter Sachverhalt wird repräsentiert durch Aussagen, durch Sätze, die in ihrer Gesamtheit das Gefüge einer Wissenschaft (oder in engerem Rahmen: einer Theorie) bilden.

Nun ist eine empirische Wissenschaft nicht einfach eine Beschreibung von Beobachtetem. Sie will darüber hinaus generalisieren auf Fälle, die noch nicht beobachtet wurden, will vorhersagen und erklären (wobei Vorhersage und Erklärung oft nur zwei Seiten derselben Medaille sind), will Zusammenhänge zwischen Variablen aufdecken, die nicht von vornherein evident sind. Dazu bedient sie sich der Logik: Aus bekannten und angenommenen (vermuteten) Sachverhalten deduziert sie neue Sätze, leitet sie neue Sachverhalte ab, die sich als beobachtbar erweisen müßten, wenn die Prämissen richtig sind. Zeigen sich diese abgeleiteten neuen Sachverhalte nicht, so muß an den Prämissen etwas falsch sein, und das Satzgebäude der wissenschaftlichen Theorie muß so lange modifiziert werden, bis der Vergleich zwischen Deduziertem (Vorhergesagtem) und beobachtetem („Daten“) befriedigend ausfällt.

## 1.1 Exkurs über Meßtheorie und Skalierung

Ein zusätzliches methodisches Problem besteht in vielen Fällen darin, wie aus Beobachtungen „Daten“ werden, die sich mit statistischen Verfahren weiterverarbeiten lassen; für die sich also Wahrscheinlichkeiten berechnen lassen. Meistens handelt es sich dabei um eine Quantifizierung der Beobachtungen, um eine Repräsentation der beobachteten Sachverhalte durch Zahlen. Mit den wissenschaftstheoretischen Problemen dieser Repräsentation beschäftigt sich die Meßtheorie, praktische Verfahren zu ihrer Lösung bezeichnet man als „Skalierungstechniken“. Auf beide Aspekte wird in dem Aufsatz von Orth (1982) in Band 3 näher eingegangen; hier sollen sie nur soweit angesprochen werden, als es im Rahmen dieses Aufsatzes erforderlich scheint.

Skalierungsverfahren bilden psychologische Variable (oder andere Sachverhalte) durch Zahlen ab, teils um sie so den Methoden der Statistik und Wahrscheinlichkeitsrechnung zugänglich zu machen, teils aber auch, um eine exaktere Festlegung und Operationalisierung dieser Variablen zu erreichen. Dabei werden beobachtbare Objekte durch Zahlen repräsentiert, beispielsweise die Intelligenz eines Menschen durch seinen Testwert, und dementsprechend Relationen zwischen Objekten durch Relationen zwischen Zahlen, beispielsweise die Relation zwischen Hubert und Jürgen, daß der eine extravertierter sei als der andere, durch die Zuordnung eines höheren Extraversionssmaßes an den einen als an den anderen: Die Relation „extravertierter als“ im Bereich der Objekte (hier: Personen) wird abgebildet durch die Relation „größer als“ im Bereich der Zahlen.

Nun kann man allerdings im Bereich der Zahlen sehr viel mehr Relationen aufstellen, als man möglicherweise im Bereich der Objekte könnte, die durch diese Zahlen abgebildet werden. Beispielsweise kann man sagen, daß 20 „doppelt so groß“ wie 10 ist, aber nicht, daß jemand mit einem „doppelt so hohen“ Extraversionstestwert auch „doppelt so extravertiert“ sei - schon einfach,

weil bei Verwendung eines anderen (aber ebenso zuverlässigen) Extraversions-tests die Relation „doppelt so hoch“ sich nicht wieder einstellen würde, wohl aber die Relation „höher als“.

Es dürfen also nur bestimmte Eigenschaften der Zahlen verwendet werden, um mit ihnen Aussagen über die von ihnen repräsentierten Objekte zu machen; im obigen Beispiel: Wahl über die Rangordnung der Zahlen, nicht aber über das Verhältnis (z.B. 1:2) zwischen ihnen. Genau genommen dürfen nur solche Eigenschaften und Relationen der Zahlen verwendet werden, die bei allen (als gleichwertig betrachteten) Repräsentationen der Objekte auftreten würden, mit anderen Worten: die invariant gegenüber zulässigen Transformationen der Zahlen sind. In diesem Sinne hat Stevens (1951) verschiedene Skalen-Niveaus definiert durch Angabe der zulässigen Transformationen der Zahlen, bei denen die durch die Zahlen repräsentierten Relationen zwischen den Objekten in der Abbildung erhalten bleiben:

Bei der sog. Nominalskala wird nur eine umkehrbar eindeutige Zuordnung der beobachteten Objekte in den Klassen gefordert. Deren Bezeichnung durch Zahlen ist willkürlich und beliebig transformierbar, soweit dabei keine Klassen zusammengelegt oder auseinandergezogen werden. Beispiele für eine „Nominalskalierung“ sind etwa Postleitzahlen, Telefonnummern oder die Kennzeichnung von Fußballspielern mit Nummern auf dem Rücken.

Bei der Ordinal- oder Rangskala wird die größer/kleiner-Relation der Zahlen mitbenutzt; zulässig sind alle Transformationen, bei denen die Rangordnung erhalten bleibt, also alle monotonen Transformationen. Dabei ist zwar sichergestellt, daß einer größeren Zahl auch eine stärkere Ausprägung der dadurch repräsentierten Eigenschaft des Objekts entspricht, aber keineswegs, ob etwa beispielsweise den gleichen Abständen  $3 - 2 = 2 - 1$  zwischen den Zahlen 1,2 und 3 auch gleiche Unterschiede in der Ausprägung der dadurch repräsentierten Eigenschaft der Objekte entsprechen.

Dies ist jedoch bei der Intervallskala sichergestellt, der nächst höheren in der Hierarchie der Skalen. Hier kann man Aussagen über Abstände zwischen den abgebildeten Objekten machen. Das Fullerton/Cattell-Prinzip beispielsweise, das Distanzen zwischen Reizen durch die relative Häufigkeit ihrer Beobachtung definiert, legt eine solche Intervallskala fest. Die aufgrund von Intervallskalen gemachten Aussagen über die Objekte sind invariant gegenüber linearen Transformationen der die Objekte repräsentierenden Zahlen.

Da der Nullpunkt einer solchen Intervall-Skala nicht festgelegt ist, können aufgrund dieser Zuordnungen immer noch keine Aussagen über Verhältnisse gemacht werden (wie z.B. „doppelt so stark ausgeprägt“ o.ä.). Dies ist erst bei der Verhältnisskala der Fall, bei der nur proportionale Transformationen zulässig sind.



Bei der absoluten Skala schließlich sind gar keine Transformationen mehr zulässig; dafür können dann auch alle Eigenschaften und Relationen der Zahlen zu Aussagen über die repräsentierten Objekte herangezogen werden.

Ziel des Forschers ist es in der Regel, ein möglichst hohes Skalenniveau zu erreichen, um in seinen Aussagen über die durch die Zahlen abgebildeten Objekte möglichst viele Relationen der Zahlen benutzen zu können. (Beispiel: Es ist informativer festzustellen, Hans sei „doppelt so schnell“ wie Karl, als nur sagen zu können, Hans sei „schneller als“ Karl. Die erste Aussage ist nach einer Messung (Skalierung) der „Schnelligkeit“ auf einer Verhältnisskala möglich, die zweite nach einer Messung auf Ordinalskalenniveau.)

Weiterhin kommt hinzu, daß es für die höheren Skalenniveaus in der Regel die effizienteren statistischen Testverfahren gibt. Die effizientesten von allen sind die sog. parametrischen Verfahren, die sich zwar mit Intervallskalenniveau begnügen, dafür aber andererseits eine ganz bestimmte Verteilungsform der Daten in der Grundgesamtheit voraussetzen, nämlich die sog. Normalverteilung. Sie gehen dabei von der Annahme aus, daß die gesamte Datenverteilung (genauer: die Verteilung der Grundgesamtheit, aus der die Daten als Stichproben entnommen sind) sich durch die Angabe zweier Parameter, nämlich des Mittelwertes ( $\mu$ ) und der Standardabweichung ( $\sigma$ ) eindeutig charakterisieren läßt. Durch diese Annahme sind die parametrischen Verfahren einerseits relativ einfach und ökonomisch durchzuführen - man braucht nur die Parameter bzw. deren Schätzungen zu betrachten - und andererseits auch sehr effizient, weil eben die gesamte Verteilungsform dabei mitbenutzt wird. Andere Verfahren - die sog. nicht-parametrischen oder verteilungs-unabhängigen - stützen sich nur auf einzelne Statistiken, die jeweils nur Teilaspekte der Verteilung repräsentieren können. Diese Verfahren sind damit zwangsläufig weniger effizient.

## 1.2 Schema des Erkenntnisgewinns in einer empirischen Wissenschaft

Die Abb. 1 (nach Coombs, Raiffa & Thrall, 1954) versucht, das Schema des Erkenntnisgewinns zu veranschaulichen: Oben findet zunächst eine Abstraktion statt, eine Reduktion der Vielfalt der beobachtbaren Welt auf einige wenige, für das jeweilige Forschungsproblem relevante Merkmale oder Variable. Wahrnehmbare Sachverhalte in der beobachtbaren Welt (oben links) werden durch Worte und Sätze oder Symbole und Formeln in der Sprache oder Theorie (oben rechts) repräsentiert oder abgebildet. Aus diesen werden (rechts im Schema) durch logische Deduktion neue Sätze oder Formeln abgeleitet. So entstehen (unten rechts) Vorhersagen oder Hypothesen über andere Sachverhalte, die ebenfalls beobachtbar werden müßten, wenn ihre Prämissen in der

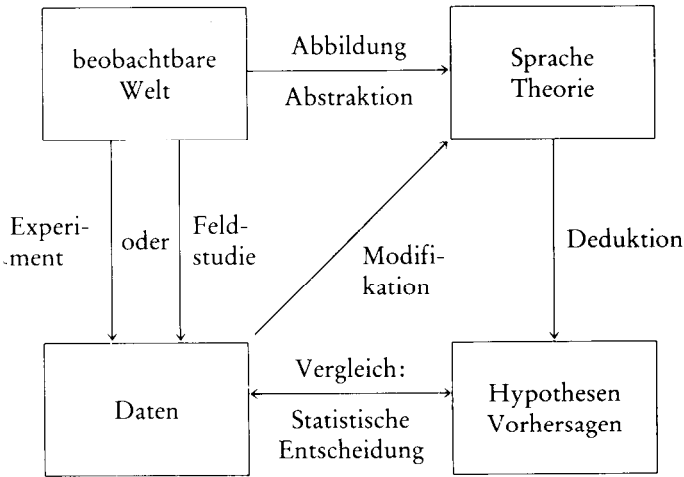


Abb. 1:

Theorie (oben rechts) richtig sind. Diese Vorhersagen oder Hypothesen müssen nun an der Realität geprüft werden. Dazu müssen in der beobachtbaren Welt (links im Schema) Bedingungen geschaffen oder aufgesucht werden, unter denen die vorhergesagten Sachverhalte (rechts im Schema) beobachtbar werden müßten: Die Herstellung solcher Bedingungen im Labor geschieht im Experiment, das Aufsuchen in der unbeeinflussten Natur in der sog. Feldstudie. Dazwischen gibt es noch das Feldexperiment. Bei diesem wird im Gegensatz zum Laborexperiment das Milieu, in dem das Experiment ablaufen soll, nicht hergestellt, sondern aufgesucht. Dann aber werden dort die relevanten Variablen wie im Experiment manipuliert, während sie in der Feldstudie selektiert werden (Bredenkamp, 1969). In all diesen Fällen macht der Forscher neue Beobachtungen, deren Ergebnisse er „Daten“ nennt (unten links im Schema). Diese Daten müssen mit den Vorhersagen (rechts unten) verglichen werden: Es muß darüber entschieden werden, ob Daten und Vorhersagen hinreichend gut übereinstimmen, so daß man annehmen kann, daß die Theorie gilt, aus der die Vorhersagen abgeleitet wurden. Bei dieser Prüfung der Übereinstimmung zwischen Daten und Vorhersagen setzt das eigentliche Thema dieses Aufsatzes ein. Zeigen die (im folgenden näher zu besprechenden) Verfahren der statistischen Entscheidungstheorie eine hinreichende Übereinstimmung an, so läßt man die Theorie, aus der die Vorhersagen abgeleitet wurden, als bestätigt gelten. Stimmen Daten und Vorhersagen nicht überein, so wird die Theorie verworfen oder zumindest modifiziert (Pfeil von unten links nach oben rechts im Schema). So kann sich aus Theorie, Vorhersagen, Daten und modifizierter Theorie ein Regelkreis bilden, von dem man sich Erkenntnisfortschritt erhofft.

Wir wollen bei alledem aber nicht außer acht lassen, daß diese Strategien eben nur von der Gemeinde der Wissenschaftler akzeptierte Konventionen sind und daß grundsätzlich auch ganz andere Strategien möglich wären. Wir werden daher zum Teil auch ihre Nachteile und alternative Lösungsansätze ansprechen.

## 2. *Klassische Statistik*

### 2.1 Vorgehensweise der klassischen Statistik

In der sog. klassischen Statistik, die im wesentlichen auf Fisher (1921, 1925) und Neyman und Pearson (1928, 1933) zurückgeführt wird, haben wir es mit folgender Vorgehensweise bei der Gewinnung oder Bestätigung neuer Erkenntnisse zu tun:

Der Forscher hat zunächst eine Hypothese, eine Aussage, die in bestimmter Weise seine Vorstellung von dem beobachteten Sachverhalt widerspiegelt oder abbildet. Es ist dabei formal zunächst unerheblich, woher diese Hypothese kommt; wichtig ist hier nur, daß sie eine Aussage über einen Sachverhalt macht, dessen Beobachtung unter den gegenwärtigen Bedingungen der Untersuchung prinzipiell möglich sein müßte.

Ein Beispiel mag das verdeutlichen: Aus einer komplexeren Persönlichkeits-theorie sei der Satz abgeleitet worden, daß Introvertierte kürzere Reaktionszeiten haben müßten als Extravertierte. Es handelt sich also um eine Hypothese über einen Zusammenhang zwischen zwei Merkmalen oder Variablen. Diese ist zunächst nur eine Vermutung, die der empirischen Bestätigung bedarf, um als Tatsache oder beobachteter Sachverhalt gelten zu können, welche die Theorie stützt.

Für die empirische Bestätigung dieser Vermutung wird man nun eine zufällige Stichprobe von Personen aus derjenigen Population suchen, für welche die Aussage gelten soll. Bei den Mitgliedern dieser Stichprobe wird man mit Verfahren, die man für geeignet hält, sowohl das Ausmaß an Extraversion/Introversion jeder Person feststellen als auch die Reaktionszeit in einer bestimmten Situation messen.

Die aus der oben angesprochenen Persönlichkeitstheorie deduzierte Hypothese für diese Versuchssituation besagt nun: Wenn allgemein (d.h. in der Grundgesamtheit oder Population, die durch diese Stichprobe repräsentiert werden soll) Introvertierte durchschnittlich kürzere Reaktionszeiten haben als Extravertierte, dann müssen auch in dieser Stichprobe die introvertierten Personen durchschnittlich kürzere Reaktionszeiten haben als die extravertierten.

Die Hypothese einer empirischen Untersuchung ist also im allgemeinen ein Satz, der eine spezifische Aussage über einen beobachtbaren Sachverhalt macht, der aus einem übergreifenden komplexeren System von Aussagen - meistens einer Theorie - abgeleitet ist, und der für die gesamte Population gelten soll. Ein Satz dieser Allgemeinheit läßt sich in der Regel empirisch nicht bestätigen. Selbst wenn sich zeigt, daß er bei den Personen der erfaßten Stichprobe zutrifft, besteht immer noch die Möglichkeit, daß er bei anderen Personen in der Population nicht mehr zutrifft; es besteht grundsätzlich die Möglichkeit, daß rein zufällig in die Stichprobe nur solche Personen hineingeraten sind, oder daß aufgrund von Meßfehlern solche Daten beobachtet werden, welche die Hypothese bestätigen, die sonst nicht richtig ist.

In der Neyman-Pearsonschen Statistik versucht man sich gegen diesen Fehler zu schützen, indem man der (oben besprochenen) eigentlichen Hypothese eine sog. Nullhypothese gegenüberstellt, welche den Inhalt der Hypothese negiert und die Abweichung der Beobachtung von der Nullhypothese auf bloßen Zufall zurückführt. Man versucht dann festzustellen, wie wahrscheinlich eine solche Zufallsbestätigung unter Annahme der Richtigkeit der Nullhypothese ist, und macht diese Wahrscheinlichkeit einer Zufallsbestätigung der Hypothese zur Grundlage ihrer Akzeptierung oder Verwerfung. Diese Nullhypothese wird verworfen und damit ihre Alternative, d.i. die ursprüngliche Forschungshypothese, angenommen, wenn die Wahrscheinlichkeit der Zufallsbestätigung der Hypothese geringer ist als ein vor der Untersuchung festgelegtes Wahrscheinlichkeitsniveau, das sog. Verlässlichkeitsniveau oder Signifikanzniveau, das mit  $\alpha$  bezeichnet wird.

Wir verwenden hier und im folgenden den Begriff der Wahrscheinlichkeit, ohne ihn näher zu definieren oder einzuführen - das wäre in diesem Rahmen nicht möglich. Im Verlauf der Geschichte sind eine Fülle von Wahrscheinlichkeitsdefinitionen vorgetragen worden; einige davon hat Nagel (1939) gesammelt. Hier soll nur soviel gesagt werden, daß auch in diesem Aufsatz (mindestens) zwei verschiedene Auffassungen von Wahrscheinlichkeit verwendet werden, je nach den Urhebern der jeweils besprochenen Verfahren. Die Wahrscheinlichkeitsauffassung der klassischen Statistik kann man als „frequentistisch“ bezeichnen: Wahrscheinlichkeit stellt einen Grenzwert einer relativen Häufigkeit dar (v. Mises, 1936; Kolmogorov, 1933).

$$p(X) = \lim_{n \rightarrow \infty} \frac{\text{Häufigkeit von X unter n Beobachtungen}}{n}$$

Schwierigkeiten bereitet dieses Konzept bei der Interpretation der Wahrscheinlichkeit einmaliger Ereignisse, für die man sich auch nicht vorstellen kann, daß man Häufigkeiten des Auftretens zählen könnte. In solchen Fällen läßt sich ein anderer Wahrscheinlichkeitsbegriff verwenden, den man als „subjektiv“ oder „subjektivistisch“ bezeichnet: Wahrscheinlichkeit als quantifizier-

tes Ausmaß der persönlichen Überzeugung dafür, daß das betreffende Ereignis eintritt (Bayes, 1763; de Finetti, 1937; Savage, 1954).

Andererseits widerstrebt ein „persönlicher“ oder „subjektiver“ Wahrscheinlichkeitsbegriff manchen, weil Wissenschaft eben etwas „Objektives“ sein soll; die Wahrscheinlichkeit soll als Eigenschaft (wie Größe und Gewicht) dem betrachteten Ereignis zukommen und nicht dem Betrachter. Hier hilft vielleicht Popper's (1976, S. 107 und 311) Interpretation der Wahrscheinlichkeit als „Propensität“, als „Maß der Verwirklichungstendenz“, die dem betrachteten Ereignis innewohnt (also nicht „subjektiv“ ist) und trotzdem auch bei neuen und einmaligen Ereignissen vorstellbar ist.

Auf Basis der historischen Wurzel, des frequentistischen Wahrscheinlichkeitsbegriffes der klassischen Statistik, bedeutet die oben angesprochene Signifikanz-Aussage also, daß solche scheinbar die Forschungshypothese bestätigende Daten bei Zutreffen der Nullhypothese (also zufällig) unter sonst gleichen Bedingungen auf lange Sicht (im Grenzwert) mit einer geringeren relativen Häufigkeit als  $\alpha$  eintreffen werden.

Um auf unser Eingangsbeispiel (Zusammenhang zwischen Extraversion und Reaktionszeit) zurückzugreifen: Die Nullhypothese würde den Zusammenhang zwischen den beiden Variablen negieren. Zeigen die Introvertierten in der Stichprobe tatsächlich kürzere Reaktionszeiten, so würde die Nullhypothese dies als Zufälligkeiten in der Zusammensetzung der Stichprobe erklären. Erweist sich die Wahrscheinlichkeit solcher Daten (nämlich entsprechend kürzerer Reaktionszeiten bei Introvertierten) unter Zufallsbedingungen als geringer als das vorher festgelegte Signifikanzniveau  $\alpha$ , so verwirft man die Zufalls-(Null-)Hypothese und nimmt einen Zusammenhang zwischen den beiden Variablen an.

Strenggenommen wird bei diesem Verfahren gar nicht die eigentliche Forschungshypothese geprüft, sondern nur die ihr gegenübergestellte Zufalls- oder Nullhypothese, und die eigentliche Forschungshypothese bleibt als Alternative übrig, wenn die Nullhypothese aufgrund zu geringer Wahrscheinlichkeit zurückgewiesen wird. Die eigentliche Forschungshypothese wird daher auch „Alternative“ oder „Alternativhypothese“ genannt. Ihre Chance, akzeptiert zu werden, hängt ab von der Wahl des Verlässlichkeitsniveaus  $\alpha$ , der Höhe der Wahrscheinlichkeit, oberhalb deren man eine Zufallsbestätigung (und damit Gültigkeit der Nullhypothese) annehmen will (allerdings auch von der Stichprobengröße). Konventionelle Grenzen für  $\alpha$  sind 0.05, 0.01 oder 0.001.

Je geringer man diese Wahrscheinlichkeit  $\alpha$  ansetzt, desto geringer wird die Chance, daß aufgrund bloßer Zufälligkeiten in der Stichprobe eine Nullhypothese verworfen und damit eine Alternative (also: Forschungshypothese) ange-

nommen wird; desto besser ist man also vor dem Irrtum geschützt, Zufälligkeiten als neue Erkenntnisse zu akzeptieren. Man spricht daher auch bei solchen niedrigen zulässigen Zufallswahrscheinlichkeiten von „hohem“ Verlässlichkeitsniveau oder Signifikanzniveau. Ein Signifikanzniveau von  $\alpha = 0.01$  bedeutet also eine höhere Verlässlichkeit (statistische Signifikanz) als ein solches von  $\alpha = 0.05$ , weil beim ersteren die Wahrscheinlichkeit des zufälligen Verwerfens einer an sich richtigen Nullhypothese geringer ist.

Nun hat diese Medaille freilich auch eine Kehrseite: Zwar wird man bei höherem Verlässlichkeitsniveau (also kleinerem Zahlenwert von  $\alpha$ ) seltener eine Forschungshypothese (Alternative) irrtümlich akzeptieren, weil nur zufällig die Verhältnisse in der Stichprobe ihr entsprachen, aber man wird auch seltener eine Forschungshypothese bestätigen, wenn sie richtig ist: Man wird bei hohem geforderten Verlässlichkeitsniveau eher noch geneigt sein, auch stärker die Forschungshypothese bestätigende Beobachtungen dem Zufall der Stichprobenauswahl zuzuschreiben. Damit ist bei hohem Verlässlichkeitsniveau gleichzeitig die Wahrscheinlichkeit der Entdeckung tatsächlich bestehender Zusammenhänge verringert.

Insgesamt gesehen, läßt das Entscheidungsmodell der klassischen Statistik also folgende Möglichkeiten zu:

Fall 1: Der in der Forschungshypothese vermutete Zusammenhang besteht nicht, und die Beobachtungen entsprechen der Zufallserwartung. Die Nullhypothese ist richtig und wird auch beibehalten. Es handelt sich also um eine richtige Entscheidung.

Fall 2: Der in der Forschungshypothese vermutete Zusammenhang besteht in Wirklichkeit (in der Grundgesamtheit) nicht; die Beobachtungen in der Stichprobe fallen aber zufällig so aus, als bestünde er. (In unserem Eingangsbeispiel zum Zusammenhang zwischen Introversion und Reaktionszeit könnte das daran liegen, daß wir zufällig viele schnelle Introvertierte und langsame Extravertierte in unserer Stichprobe haben, während in der Grundgesamtheit kein solcher Zusammenhang besteht.) Verwerfen wir in solch einem Falle die Nullhypothese und nehmen wir die Alternative an, so begehen wir einen Fehler, und zwar den sog.  $\alpha$ -Fehler oder Fehler I. Art. Die Wahrscheinlichkeit seines Auftretens ist durch die Höhe des gewählten Verlässlichkeitsniveaus  $\alpha$  begrenzt: Man wird aber diesen Fehler bei einem festgelegten Verlässlichkeitsniveau  $\alpha$  nur mit der Wahrscheinlichkeit  $\alpha$ , oder - frequentistisch gedeutet - auf lange Sicht relativ nur so häufig begehen, wie dieses  $\alpha$  angibt.

Fall 3: Der in der Forschungshypothese vermutete Zusammenhang besteht tatsächlich, und die Beobachtungen in der empirischen Untersuchung

bestätigen dies hinreichend stark, so daß die Wahrscheinlichkeit eines zufälligen Zustandekommens dieser Beobachtungen nicht höher als das gewählte Verlässlichkeitsniveau  $\alpha$  ist. Die Nullhypothese wird daher verworfen, und der in der Alternative (Forschungshypothese) vermutete Zusammenhang als empirisch bestätigt angenommen. Wie im Fall 1 handelt es sich um eine richtige Entscheidung, diesmal zugunsten der Forschungshypothese.

Fall 4: Der in der Forschungshypothese vermutete Zusammenhang besteht tatsächlich, aber die Beobachtungen zeigen dies nicht deutlich genug oder das Verlässlichkeitsniveau ist so hoch angesetzt ( $\alpha$  so klein), daß die Ergebnisse nicht mit hinreichender Sicherheit vom Zufall unterschieden werden können, also ihr zufälliges Zustandekommen nicht hinreichend wenig wahrscheinlich ist (ihre Zufallswahrscheinlichkeit größer als das gewählte  $\alpha$  ist), so daß die Nullhypothese beibehalten und die (an sich richtige) Alternative nicht bestätigt werden kann. Es wird also ein in der Grundgesamtheit tatsächlich bestehender Zusammenhang nicht erkannt, sondern die Nullhypothese fälschlich beibehalten. Auch hier begehen wir einen Fehler, und zwar den  $\beta$ -Fehler oder Fehler II. Art. Während die Wahrscheinlichkeit des  $\alpha$ -Fehlers oder Fehlers erster Art nach oben durch die Wahl des Verlässlichkeitsniveaus  $\alpha$  begrenzt war, können wir über die Wahrscheinlichkeit des Auftretens eines Fehlers II. Art wenig aussagen. Wohl bezeichnet man seine Wahrscheinlichkeit (analog zu der des  $\alpha$ -Fehlers) mit  $\beta$ , aber meistens fehlt jede Angabe darüber, wie groß dieses  $\beta$  sein könnte, wie wir im folgenden sehen werden. Trotzdem gibt es in gewissen Grenzen Möglichkeiten, seine Höhe zu beeinflussen:

- (a) Je höher man das Verlässlichkeitsniveau und je kleiner man damit die Wahrscheinlichkeit  $\alpha$  wählt, desto größer wird unter sonst gleichen Bedingungen die Wahrscheinlichkeit  $\beta$  ausfallen.
- (b) Je größer man die beobachtete Stichprobe wählt, desto deutlicher werden sich auch in ihr systematische Zusammenhänge zeigen, die in der Grundgesamtheit vorhanden sind, desto größer wird also die Chance, diese zu entdecken, und desto geringer wird damit die Wahrscheinlichkeit  $\beta$  eines Fehlers II. Art.
- (c) Je besser ein statistischer Test die in den Beobachtungen (Daten) enthaltene Information ausnutzt, desto eher wird es mit seiner Hilfe gelingen, unter sonst gleichen Bedingungen eine falsche Nullhypothese zurückzuweisen. Man nennt diese Eigenschaft eines statistischen Tests (nämlich seine Fähigkeit, unter Berücksichtigung der gegebenen Information falsche Nullhypothesen zu verworfen) daher auch seine „Teststärke“ und gibt sie als  $1-\beta$  an. Wenn man die Wahl hat, wird man also einen in diesem Sinne „stärkeren“ Test einem schwächeren vorziehen. Allerdings stellen

stärkere Tests in der Regel auch strengere Anforderungen an die Qualität des Datenmaterials. Die stärksten Tests sind in jedem Falle die sog. parametrischen Tests, die normalverteilte Daten voraussetzen und von der Annahme ausgehen, daß hinter den beobachteten Daten normalverteilte Grundgesamtheiten stehen mit genau den Parametern, die man aus den Kennwerten (Mittelwerte, Varianzen, Kovarianzen) der Stichprobe für sie geschätzt hat. Kann man die Annahme der Normalverteilung der Daten nicht aufrechterhalten, so kann man streng genommen nur einen Test geringerer Teststärke verwenden.

- (d) Je weiter der unter der Alternativhypothese angenommene Zustand von dem unter der Nullhypothese angenommenen entfernt ist, je stärker also beispielsweise der tatsächliche Unterschied zwischen verschiedenen behandelten Gruppen ist (die unter Annahme der Nullhypothese Stichproben aus der gleichen Grundgesamtheit sein müßten), desto leichter wird ein Test diesen Unterschied entdecken, also die Nullhypothese verwerfen können. Diese Abhängigkeit der Wahrscheinlichkeit  $\beta$  eines Fehlers II. Art von dem tatsächlichen Zustand der Natur wird Operationscharakteristik des Tests genannt. Anders ausgedrückt: Die Wahrscheinlichkeit  $(1 - \beta)$  der Entdeckung eines Unterschiedes zwischen den Gruppen von der Größe des tatsächlich bestehenden Unterschiedes bezeichnet man als Gütefunktion des Tests. Die Operationscharakteristik und ihr Komplement, die Gütefunktion, verlaufen bei allen sinnvollen Tests so, daß mit zunehmendem Ausmaß des Unterschiedes auch die Wahrscheinlichkeit seiner „Entdeckung“ oder statistischen Sicherung durch den Test wächst.
- (e) Die statistische Sicherung einer Hypothese gegen die Null- oder Zufallshypothese geschieht bei den klassischen Signifikanztests auf Basis der sog. Fehlervarianz, der nicht mehr weiter erklärbaren Verschiedenheit zwischen Elementen innerhalb der gleichen Stichprobe. Die Unterschiede zwischen verschiedenen behandelten oder ausgewählten Gruppen (über welche die Alternativhypothese eine Aussage macht) werden im Prinzip verglichen mit den Unterschieden zwischen den Gruppen, die man erwarten würde, wenn die Gruppen Stichproben aus der gleichen Grundgesamtheit wären (wie es die Nullhypothese annimmt). Diese Verschiedenheit zwischen den Gruppen, die man rein zufällig bei Zutreffen der Nullhypothese erwarten würde, ist aber in ihrem Ausmaß abhängig von der Verschiedenheit der Elemente innerhalb der gleichen Gruppe, der sog. Fehlervarianz. Gelänge es uns nun also, diese Fehlervarianz und damit die bei Zutreffen der Nullhypothese erwartete Verschiedenheit oder Varianz zwischen den Gruppen zu verringern, so werden die tatsächlich beobachteten Unterschiede zwi-



schen den Gruppen leichter die Signifikanzgrenze erreichen. Der statistische Test ist also um so effizienter, je geringer die Fehlervarianz, die Verschiedenheit innerhalb der Gruppe ist. Dies läßt sich praktisch erreichen (beispielsweise bei varianzanalytischen Versuchsplänen) durch Einführung von Kovariaten (Kovarianzanalyse, Regressionskorrektur der abhängigen Variablen) oder durch Einführung von zusätzlichen Faktoren (unabhängigen Variablen) zur Blockbildung, oder - wo das möglich ist - durch Meßwiederholung an den gleichen Vpn. (In den letztgenannten Fällen wird die Fehlervarianz um die Varianz „zwischen den Blöcken“ oder „zwischen den Vpn“ reduziert, und damit die Effizienz des Tests erhöht.)

So können zwar eine Reihe von Aussagen über das Verhalten der Wahrscheinlichkeit  $\beta$  des Fehlers II. Art gemacht werden; ihr numerischer Wert bleibt aber in der Regel unbekannt. Das liegt daran, daß die Forschungshypothesen, die Alternativhypothesen der statistischen Tests, in den meisten Fällen nicht so präzise formuliert sind, daß man eine Wahrscheinlichkeitsverteilung der Beobachtungen oder Daten aus ihnen herleiten könnte, wie dies bei der Nullhypothese der Fall ist. Etwas überspitzt kann man sogar sagen, daß die Nullhypothese aufgestellt wird, gerade weil man für sie eine Verteilung der Daten vorhersagen kann.

Mit diesem Problem hat sich Cohen (1962, 1969, 1977) auseinandergesetzt, im deutschen Sprachraum auch Bredenkamp (1972), Henning & Muthig (1979) sowie Witte (1980). Cohen (1977) hat praktisch anwendbare Formeln und Tabellen für eine Vielzahl konkreter Fälle zusammengestellt, die z.T. bei Henning & Muthig (1979) referiert werden.

In die Überlegungen zur Power-Analyse gehen im wesentlichen vier Variablen ein: die Effektgröße (z.B. die Größe eines Mittelwertunterschieds), die Stichprobengröße, das Verlässlichkeitsniveau  $\alpha$  und die Teststärke (Power)  $1 - \beta$ . Sind drei von diesen vier Variablen bekannt, so kann man die vierte aus ihnen berechnen (Näheres s. Henning & Muthig, 1979, S. 228ff.; oder Cohen, 1977).

Gehen wir zurück auf unser Eingangsbeispiel:

Die Forschungshypothese besagte, daß Introvertierte schneller reagieren als Extravertierte - sie sagt aber nicht, um wieviel schneller. Die entsprechende Nullhypothese besagt demgegenüber, daß Intro- und Extravertierte gleich schnell reagieren. Für diese Nullhypothese läßt sich ohne weiteres eine Verteilung der Reaktionszeiten ableiten: Die voraussichtliche Differenz der Stichproben-Mittelwerte zwischen Introvertierten und Extravertierten muß sich in dem Rahmen bewegen, der als Standardfehler des arithmetischen Mittels durch die Gruppengrößen  $n_1$  und  $n_2$  und durch die Schätzung der Varianz der Meß-

werte gegeben ist. In diesem Rahmen der Verteilung von Differenzen von Mittelwerten von Stichproben aus gleichen Grundgesamtheiten läßt sich die Wahrscheinlichkeit der tatsächlich beobachteten Mittelwertsdifferenz bestimmen, so daß diese mit dem gewählten  $\alpha$  verglichen und eine Entscheidung über die Nullhypothese gefällt werden kann.

Nicht so unter der Alternativhypothese. Wollte man für sie eine vergleichbare Wahrscheinlichkeitsverteilung für mögliche auftretende Mittelwertsdifferenzen aufstellen, so müßte man genau festlegen, wie groß der Mittelwertsunterschied in der Population sein soll. Dann könnte man unter Annahme dieses Grundgesamtheitsparameters - wie bei der Nullhypothese, wo diese mit Null angenommen wurde - eine Wahrscheinlichkeitsverteilung für die beobachtbaren Stichprobenmittelwertsdifferenzen aufstellen, und damit die Wahrscheinlichkeit  $\beta$  eines Fehlers II. Art unter gegebenen Bedingungen festlegen. Wir hätten es hier mit der bereits oben angesprochenen Operationscharakteristik zu tun. Aber in den meisten Fällen sind Forschungshypothesen nicht so spezifisch, daß sie (in unserem Beispiel) eine genaue Differenz festlegen; sie begnügen sich mit einem ordinalen „mehr“ oder „weniger“, für das die Aufstellung solcher Wahrscheinlichkeitsverteilungen (und damit die Festlegung von  $\beta$ ) nicht möglich ist.

Die Wahrscheinlichkeit, einen Fehler II. Art zu begehen, also das  $\beta$ -Risiko, hängt ab von der Stärke des tatsächlichen Zusammenhanges, im Beispiel also von der Größe der Mittelwertdifferenz: Je größer dieser Unterschied, oder allgemeiner: je stärker ein Zusammenhang ist, desto größer ist auch die Wahrscheinlichkeit  $1-\beta$ , daß dieser Zusammenhang in einer Stichprobe entdeckt und „signifikant“ wird. Charakterisieren wir den tatsächlichen Zustand der Grundgesamtheit - im Beispiel also die Differenz zwischen den mittleren Reaktionszeiten der Intro- und Extravertierten - durch einen Parameter  $\Theta$ , so können wir die funktionale Abhängigkeit zwischen diesem Parameter  $\Theta$  und der Wahrscheinlichkeit  $(1-\beta)$  der Entdeckung dieses Zustandes durch eine Kurve beschreiben, eben die oben besprochene Gütefunktion des zur Entscheidung verwendeten Tests.

Obwohl allgemein zu erwarten ist (oder zumindest von einem brauchbaren statistischen Test verlangt werden sollte), daß die Entdeckungswahrscheinlichkeit  $(1-\beta)$  mit Wachsändern  $\Theta$  (hier: Mittelwertsunterschied in der Grundgesamtheit) wächst, so kann der *Verlauf* dieser Kurve bei verschiedenen statistischen Tests doch sehr unterschiedlich sein. Ein Test ist allgemein um so besser zur Entdeckung von Zusammenhängen in Grundgesamtheiten geeignet, je steiler die Operationscharakteristik ansteigt. Man bezeichnet Tests mit steilem Anstieg der Operationscharakteristik als „effizienter“, weil man im Durchschnitt mit kleineren Stichproben auskommt als bei weniger effizienten Tests (mit flacherem Anstieg der Operationscharakteristik). Unter den herkömmlichen statistischen Signifikanztests, die man in jedem Lehrbuch der

Quantitativen Methoden der Psychologie oder Statistik für Psychologen usw. abgehandelt findet, sind die effizientesten in dem obigen Sinne die sog. „parametrischen“ Tests. Diese setzen voraus, daß die Daten in der Grundgesamtheit normalverteilt sind und daher in ihrer Gesamtheit stellvertretend durch die Parameter Mittelwert ( $\mu$ ) und Standardabweichung ( $\sigma$ ) repräsentiert werden können, die aus der Stichprobe geschätzt werden. Die klassischen parametrischen Tests wie beispielsweise t- und F-Test (einschließlich der Varianzanalyse) verdanken ihre hohe Effizienz letztlich dem Umstand, daß sie nicht nur Einzelcharakteristika der Stichproben auf ihre Zufallswahrscheinlichkeit prüfen, sondern im Grunde genommen die ganze Verteilungsform, repräsentiert durch die Parameter  $\mu$  und  $\sigma$ .

## 2.2 Eigenschaften klassischer Tests

Aus dem oben Gesagten ergeben sich eine Reihe von Anforderungen oder Qualitätsansprüchen, die an Signifikanztests gestellt werden, mit denen Entscheidungen über Forschungshypothesen nach dem obigen Prinzip gefällt werden sollen. Die wichtigsten dieser Forderungen sind:

1. Konsistenz: Wenn die Nullhypothese falsch ist, soll mit wachsender Stichprobengröße ( $n \rightarrow \infty$ ) die Teststärke ( $1 - \beta$ ) gegen 1 gehen.
2. Dominanz: Bei festgelegtem Verlässlichkeitsniveau  $\alpha$  und festgelegter Stichprobengröße  $n$  soll es keinen anderen Test geben, dessen Teststärke ( $1 - \beta$ ) größer ist. Ein Test, der dieses Kriterium erfüllt, wird dann auch als „bester Test“ bezeichnet.
3. Unverfälschtheit oder Erwartungstreue: Die in dem Test verwendeten Stichprobenfunktionen, z.B. Schätzfunktionen für Parameter der Grundgesamtheit, sollen erwartungstreu sein, d.h., ihr Erwartungswert soll der „wahre“ Parameter der Grundgesamtheit sein.
4. Suffizienz: Die in dem Test verwendeten Stichprobenfunktionen sollen die gesamte für die Entscheidung relevante Information der Stichprobe ausschöpfen.

Ausführlichere Informationen über Eigenschaften von (und Anforderungen an) klassische Signifikanztests findet der Leser beispielsweise in Fisz (1970), Kap. 11.

Ein Test, der eher zur Beibehaltung als zur Ablehnung der Nullhypothese führt (also eher zum Fehler II. Art neigt), wird auch als „konservativ“ bezeichnet; dementsprechend ein Test, der in die andere Richtung neigt, als „radikal“.

Die oben genannten Eigenschaften statistischer Tests sollen helfen, in einer Entscheidungssituation eine Wahl unter ihnen zu treffen. Nun wird man in den meisten Fällen geneigt sein, unter den bei dem gegebenen Skalenniveau der Daten zulässigen Tests den effizientesten auszuwählen, bei dem das Beta-Risiko am geringsten ist. Das muß aber nicht unbedingt die klügste Wahl sein; wenn bei einem massiven Effekt in den Daten ein weniger effizienter, aber in der Handhabung einfacherer (Rechen-Aufwand) Test bereits ausreicht, um die Nullhypothese zu verwerfen, so geht man ja kein Beta-Risiko mehr ein. Hier kann man also auch bei der Datenauswertung ökonomische Gesichtspunkte berücksichtigen. Dies darf nun aber nicht dazu führen, daß man der Reihe nach alle möglichen Tests ausprobiert, bis einer die ersehnte „Signifikanz“ erbringt - da die Tests zum Teil auf unterschiedlichen Stichprobenfunktionen aufbauen, bedeutet dieses „snooping in the methods“ letztendlich eine Aufweichung des Verlässlichkeitsniveaus. Wenn die Wahrscheinlichkeit, zufällig „signifikant“ zu werden, für jeden einzelnen Test kleiner oder gleich  $\alpha$  ist, so kann die Wahrscheinlichkeit, daß der eine oder andere „signifikant“ wird, durchaus größer sein. Im Grunde sollte die Wahl des Signifikanztests ebenso wie die des Verlässlichkeitsniveaus *vor* Durchführung der Untersuchung geschehen.

Logisch gesehen, folgt die Neyman-Pearsonsche Entscheidungstechnik dem Schlußprinzip des Modus tollens: Es wird eine Annahme  $H_0$  über eine bestimmte Verteilungsform mit bestimmten Parametern getroffen. Aus dieser folgt, daß die Daten  $D_0$  einer beobachteten Stichprobe sich in bestimmter Weise verteilen müssen. Tun sie dies nicht (oder genauer: ist die Wahrscheinlichkeit dafür zu gering, nämlich kleiner als  $\alpha$ ), so entscheidet man, daß  $H_0$  falsch sei, und damit ihr Gegenteil  $H_A$  richtig. Etwas formaler ausgedrückt:

Aus  $H_0$  folgt  $D_0$   
 nun aber nicht  $D_0$

---

also nicht  $H_0$

Da aber nicht  $H_0$ , folgt  $H_A$  (quoniam tertium non datur)

Dieses sieht wie ein logisch einwandfreier Schluß aus, ist aber streng genommen keiner, denn wir können nicht mit endgültiger Sicherheit sagen „nun aber nicht  $D_0$ “, sondern nur: „nun aber ist die Wahrscheinlichkeit der Daten unter Annahme des Zutreffens von  $H_0$  kleiner als  $\alpha$ “.

Der Gewinn neuer Erkenntnisse nach dem Neyman-Pearsonschen Signifikanztest erfolgt also über die Falsifikation, das Verwerfen einer Nullhypothese. Kann die Nullhypothese dagegen nicht verworfen werden, so kann daraus nicht logisch geschlossen werden, daß sie „wahr“ ist, denn man kann nicht schließen:

Aus  $H_0$  folgt  $D_0$   
 nun aber  $D_0$

---

also  $H_0$ ,

da dies *kein* zulässiger deduktiver Schluß ist: Die Daten  $D_0$  können auch aus ganz anderen Gründen als wegen Gültigkeit von  $H_0$  aufgetreten sein.

Es gibt also keine formale Basis, die Nullhypothese aufgrund klassischer statistischer Tests anzunehmen, denn der obige induktive Schluß ist im Rahmen der klassischen deduktiven Logik nicht zu rechtfertigen. Andererseits kommt man in einer empirischen Wissenschaft mit der rein deduktiven Logik nicht aus; sie ist letzten Endes zirkulär und tautologisch und ermöglicht keine neuen Erkenntnisse. Carnap (1959) hat die Möglichkeit einer induktiven Logik aufgezeigt; wir verweisen in diesem Zusammenhang auf Carnap & Stegmüller (1959) und Kutschers (1972).

Dies ist besonders problematisch in solchen Fällen, in denen die Nullhypothese selbst die eigentliche Forschungshypothese ist, beispielsweise wenn ein Forscher zeigen will, daß bestimmte Daten nur zufällig von den Vorhersagen seines Modells oder von einer bestimmten Verteilungsform (z.B. bei der Prüfung auf Normalverteilung als Voraussetzung für die Anwendung parametrischer Testverfahren) abweichen: Ganz abgesehen davon, daß man hierbei einen  $\beta$ -Fehler mit unbekannter Wahrscheinlichkeit in Kauf nimmt, hat man zusätzlich den Nachteil, die Beibehaltung der (Null-)Hypothese auf einem logisch unzulässigen Schlußprinzip aufzubauen. Die sich daraus für die praktische Forschungsarbeit ergebenden Probleme diskutieren beispielsweise Cook, Gruder, Henningan & Flay (1979) an einem Beispiel aus der Sozialpsychologie.

Die oben angesprochenen logischen Schlußfiguren sind an sich deterministische Prinzipien. Sie werden auf die probabilistischen, d.h. Zufallsschwankungen unterworfenen Daten einer empirischen Wissenschaft angewendet unter Annahme des Verlässlichkeitsniveaus als kritischer Grenze, das in der Statistik eine ähnliche Rolle spielt wie der Begriff der Schwelle in der Psychophysik: Was eine höhere Zufallswahrscheinlichkeit hat als eine gewisse vorgegebene  $\alpha$ , gilt nicht mehr als „wahr“.

## 2.3 Zur Frage der Stichprobengröße

Bei Vorliegen einer präzisen Alternativhypothese  $H_1$ , z.B.  $\Theta = \Theta_1$ , ist es für die klassischen Testverfahren möglich, den für eine Entscheidung bei gegebenem  $\alpha$  und  $\beta$  erforderlichen Stichprobenumfang zu berechnen.

Da dies in der Beratungspraxis des Verfassers eine häufig gestellte Frage ist, auf welche die herkömmlichen Statistik-Lehrbücher aber nicht ausreichend antworten, wollen wir kurz darauf eingehen. Einfachheitshalber und zum leichteren Verständnis führen wir unsere Überlegungen am Beispiel eines einfachen Mittelwertsvergleichs von Stichproben aus normalverteilten Grundgesamtheiten mit gleicher Varianz durch:

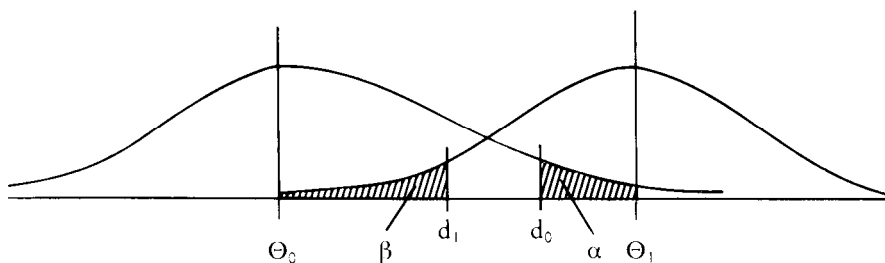


Abb. 2a:

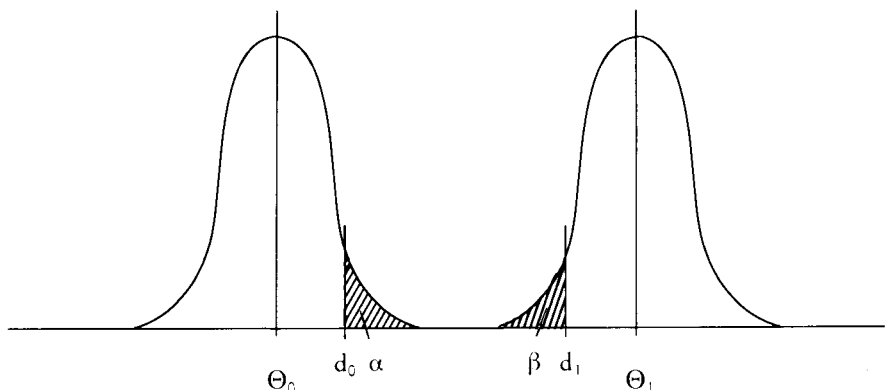


Abb. 2b:

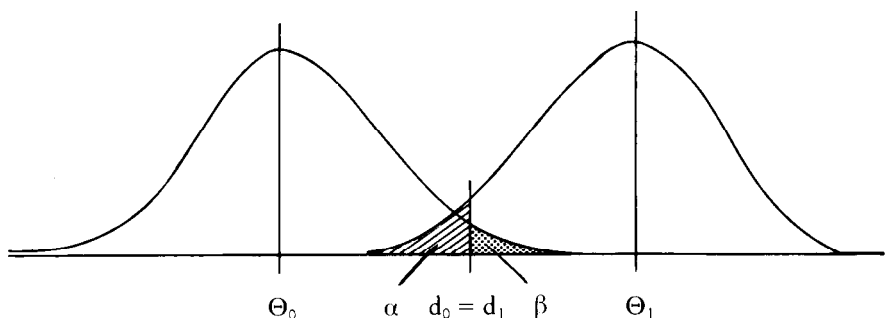


Abb. 2c:

Die konkurrierenden Hypothesen seien

$$H_0 : \Theta = \Theta_0,$$

$$H_1 : \Theta = \Theta_1,$$

in beiden Fällen sei die Standardabweichung der Grundgesamtheit gleich  $\sigma$ . Gesucht ist die Stichprobengröße  $N$ , aufgrund derer bei gegebenen Fehlentscheidungsrisiken  $\alpha$  und  $\beta$  zwischen  $H_0$  und  $H_1$  entschieden werden kann. Aufgrund der Normalverteilung der Grundgesamtheit und/oder nach dem zentralen Grenzwertsatz sind auch die Mittelwerte von Stichproben der Größe  $N$  normalverteilt, und zwar mit einer Standardabweichung (= Standardfehler des arithmetischen Mittels) von  $\sigma / \sqrt{N}$ .

Die Abb. (2) veranschaulicht die Verteilungen möglicher Stichprobenmittelwerte um die hypothetischen Grundgesamtheitsmittelwerte  $\Theta_0$  und  $\Theta_1$  bei hinreichend großen Stichproben.

Die Hypothese  $H_0 : \Theta = \Theta_0$  soll verworfen werden, wenn ein Stichprobenmittelwert gefunden wird, für den die Wahrscheinlichkeit diesen oder einen größeren zu finden bei Zutreffen von  $H_0$  kleiner (oder gleich)  $\alpha$  ist. Der Anteil  $\alpha$  ist rechts von der Verteilung möglicher Stichprobenmittelwerte aus der Grundgesamtheit mit  $\Theta = \Theta_0$  bei  $d_0$  abgeschnitten; nach diesem Kriterium ( $\alpha$ ) wird man  $H_0$  also verwerfen, wenn ein Stichprobenmittelwert rechts von (größer als)  $d_0$  auftritt. Nach dem anderen Kriterium,  $\beta$ , soll  $H_1$  verworfen werden, wenn der beobachtete Stichprobenmittelwert in der Verteilung um  $\Theta_1$  eine Wahrscheinlichkeit kleiner oder gleich  $\beta$  hat; dieser Anteil der Verteilung um  $\Theta_1$  ist links von dieser Verteilung bei  $d_1$  abgeschnitten.

In Abb. (2a) ist die Überlappung der beiden Verteilungen so groß, daß  $d_1$  links von  $d_0$  liegt; würde man hier beim Auftreten eines Stichprobenmittelwertes größer als  $d_0$  zugunsten von  $H_1$  entscheiden und bei Mittelwerten unter  $d_0$  zugunsten von  $H_0$ , so würde man zwar das gewählte  $\alpha$ -Risiko einhalten, aber eine größere Wahrscheinlichkeit eines Fehlers der II. Art als  $\beta$  in Kauf nehmen, nämlich so groß, wie der Anteil der Verteilung um  $\Theta_1$  von  $d_0$  angibt.

Würde man andererseits  $d_1$  in Abb. (2a) zum kritischen Wert machen und bei Stichprobenmittelwerten unter  $d_1$  zugunsten von  $H_0$  und bei Stichprobenmittelwerten oberhalb von  $d_1$  zugunsten von  $H_1$  entscheiden, so hielte man zwar das gewählte  $\beta$ -Risiko ein, nähme aber eine Wahrscheinlichkeit des Fehlers I. Art in Kauf, die größer als  $\alpha$  ist.

In Abb. (2b) haben die Verteilungen der Stichprobenmittelwerte um  $\Theta_0$  und  $\Theta_1$  kleinere Streuungen; die Beziehung zwischen  $d_0$  und  $d_1$  kehrt sich dadurch um.

Jetzt liegt die  $\alpha$  repräsentierende Fläche oberhalb von  $d_0$ . Wählen wir jetzt  $d_0$  zum Kriterium, so haben wir die gewünschte Wahrscheinlichkeit von  $\alpha$  für

den Fehler I. Art, aber die Wahrscheinlichkeit für den Fehler II. Art ist kleiner als  $\beta$ .

Entsprechend entscheiden wir beim Kriterium  $d_1$  zwar mit der gewünschten Wahrscheinlichkeit von  $\beta$  für den Fehler II. Art, aber mit einem kleineren  $\alpha$ -Risiko.

Mit genau den gewählten Irrtumswahrscheinlichkeiten  $\alpha$  und  $\beta$  für den Fehler I., bzw. II. Art würden wir entscheiden, wenn  $d_0$  und  $d_1$  in einem Punkt zusammenfallen. Da  $\Theta_0$ ,  $\Theta_1$  und  $\sigma$  festliegen, die Standardabweichungen der Stichprobenmittelwerte

$$\sigma(M) = \sigma / \sqrt{N}$$

aber von der Stichprobengröße  $N$  abhängt, können wir nun  $N$  so wählen, daß  $d_0$  und  $d_1$  gleich werden:

$$d_0 = d_1 = d.$$

Dieser Wert  $d$  hat die Eigenschaft, daß rechts von ihm der Anteil  $\alpha$  der Verteilung der Stichprobenmittelwerte um  $\Theta_0$  liegen, und links von ihm der Anteil  $\beta$  der Verteilung der Stichprobenmittelwerte um  $\Theta_1$  liegt. Diesen Wert  $d$  verwandeln wir nun in beiden Verteilungen in einen Standard-z-Wert:

$$z_0 = \frac{d - \Theta_0}{\frac{\sigma}{\sqrt{N}}} = \frac{(d - \Theta_0) \sqrt{N}}{\sigma} \quad \text{und}$$

$$z_1 = \frac{d - \Theta_1}{\frac{\sigma}{\sqrt{N}}} = \frac{(d - \Theta_1) \sqrt{N}}{\sigma}$$

(wobei  $z_0 > 0$ . und  $z_1 < 0$  ist).

Die Werte  $z_0$  und  $z_1$  sind aber diejenigen Abszissenwerte von Standardnormalverteilungen, bei denen die Anteile  $\alpha$  (rechts von  $z_0$ ) bzw.  $\beta$  (links von  $z_1$ ) von der Verteilung abgeschnitten werden. Sie können also aus der Standard-Normalverteilungs-(z-)Tab bei den gewählten Werten von  $\alpha$  und  $\beta$  abgelesen und in die obigen Gleichungen eingesetzt werden.

Subtrahieren wir die beiden Gleichungen voneinander, so erhalten wir

$$z_0 - z_1 = \frac{\sqrt{N}}{\sigma} (\Theta_1 - \Theta_0) \quad \text{und damit}$$

$$N = \left( \frac{z_0 - z_1}{\Theta_1 - \Theta_0} \right)^2 \cdot \sigma^2$$



für die gesuchte Stichprobengröße. (Hierbei ist zu beachten, daß  $z_1 < 0$  war, somit im Zähler des Bruches mit  $|z_0| + |z_1|$  praktisch die Summe der zu  $\alpha$  und  $\beta$  gehörigen z-Werte steht.)

Diese Formel für den Stichprobenumfang hat den didaktischen Vorteil, relativ leicht verständlich und in ihrer Entwicklung leicht nachvollziehbar zu sein - in Wirklichkeit geht sie von vereinfachenden Annahmen aus (Normalverteilung) und stellt für die Praxis nur eine grobe Abschätzung des benötigten Stichprobenumfanges dar. Für eine ausführliche Beschäftigung mit dem Problem sei auf Cohen (1977, S. 456) verwiesen; die wesentlichen Grundgedanken daraus referierten auch Henning & Muthig (1979) und Witte (1980, S. 132, 148, 155ff., 168, 175). Im wesentlichen geht es immer darum, daß von vier Variablen drei gegeben sein müssen und die vierte daraus erschlossen werden kann :

- (1) Effektgröße (bei uns  $(\Theta_1 - \Theta_2)$ ; hierzu später mehr)
- (2) Verlässlichkeitsniveau  $\alpha$
- (3) Teststärke  $(1 - \beta)$
- (4) Stichprobengröße (N)

## 2.4 Zur Effektstärke

Wie wir oben gesehen haben, gehen in die Entscheidungstechnik der klassischen Statistik zum Verwerfen und Akzeptieren von Hypothesen kaum Überlegungen ein, welche die Größe oder Stärke eines zu entdeckenden Effekts berücksichtigen. Lediglich in unserem Exkurs über die für eine Entscheidung erforderliche Stichprobengröße ging der Abstand der konkurrierenden Hypothesen  $(\Theta_1 - \Theta_0)$  mit ein.

Im Grunde kann man auf der Grundlage der klassischen Statistik jeden beliebig kleinen Unterschied durch Wahl einer entsprechend großen Stichprobe „signifikant machen“. Dies erscheint von der Sache her nicht sinnvoll, und man hat daher versucht - auch im Rahmen der klassischen Statistik-Effektstärke-Indices einzuführen, die von dem Signifikanz-Konzept unabhängig sind. Grundgedanke ist dabei - besonders im Rahmen der parametrischen Statistik - das Verhältnis von Varianzanteilen nach dem Muster des Determinationskoeffizienten, der bei einer einfachen linearen Korrelation angibt, welcher Anteil der Varianz einer Variablen durch die einer anderen, mit ihr korrelierten Variablen festgelegt ist:

$$\sigma_y^2 = \sigma_y^2 r_{xy}^2 + \sigma_y^2 (1 - r_{xy}^2)$$

Solche Maße haben den Vorteil der direkten, anschaulichen Interpretierbarkeit für die Breite von Vertrauensintervallen: Will ich beispielsweise einen Schul-

erfolg aufgrund einer Testleistung vorhersagen und kann ich „vorher“, also ohne Kenntnis der Testleistung nur aufgrund der Verteilung der Schulleistungen sagen, daß die voraussichtliche Schulleistung mit einer Wahrscheinlichkeit von 0.9 in einem Bereich von  $y_1$  bis  $y_2$  liegt, so läßt sich dieser Vertrauensbereich der Schätzung bei Kenntnis der Testleistung auf  $\sqrt{(1 - r_{ST}^2)} (y_2 - y_1)$  verkleinern - „schrumpfen“, daher auch der Name „Schrumpfungskoeffizient“ für  $\sqrt{(1 - r^2)}$ .

Einen derart anschaulichen Gebrauch kann man von den für varianzanalytische Versuchspläne eingeführten Effektivitätsmaßen nicht machen, aber das dahinterstehende Prinzip ist das gleiche wie beim Determinationskoeffizienten  $r^2$ : Man will angeben, zu welchem Anteil die Varianz der abhängigen Variablen durch die unabhängige(n) determiniert ist.

In direkter Anknüpfung an den Determinationskoeffizienten der Korrelationsrechnung hat hierfür beispielsweise Bredenkamp (1970) für den Fall des Vergleichs zweier unabhängiger Stichproben - also den t-Test - den quadrierten punkt-biserialen Korrelationskoeffizienten vorgeschlagen; dem entspricht etwa Hays (1963)  $\omega^2$ , das Verhältnis zwischen den (erwartungstreuen) Schätzungen der Varianz „zwischen“ und „innerhalb“ der Stichproben. Das in der Varianzanalyse häufig verwendete Effektmaß  $\eta^2$  oder  $\omega^2$  (Hays, 1963) ist das Verhältnis zwischen der Varianzschätzung „zwischen“ und der totalen Varianzschätzung und entspricht damit am ehesten dem Determinationskoeffizienten aus der Korrelationsrechnung. Cohens (1977, S. 274ff.) Effektmaß  $f$  setzt anstelle der Varianzschätzungen die Standardabweichung der Mittelwerte ins Verhältnis zur Standardabweichung innerhalb der Gruppen. Zwischen  $\eta^2$  und Cohens  $f$  besteht die Beziehung  $\eta^2 = f^2/(1 + f^2)$ . (Näheres hierzu s. Bredenkamp (1970), Bredenkamp (1972, S. 47ff.) oder Witte (1980, S. 151 ff.).)

Wegen der oben angesprochenen Möglichkeit (oder: Gefahr), daß auch winzige Effekte durch hinreichend große Stichproben statistisch „signifikant“ werden können, sollten bei der Beurteilung von Hypothesen immer solche Effektmaße mit betrachtet werden.

## 2.5 Zusammenfassung des klassischen Signifikanztests

Wir können nun die Vorgehensweise des klassischen Neyman-Pearsonschen Signifikanztests in folgender Weise zusammenfassen:

Einer theoretisch abgeleiteten Forschungshypothese  $H_A$  wird eine Nullhypothese  $H_0$  gegenübergestellt, welche die Forschungshypothese (Alternative)  $H_A$  negiert und die in der Untersuchung gemachten Beobachtungen (Daten) als zufällig um den in  $H_0$  angenommenen Parameter variierend, d.h. ohne Wir-

kung des in  $H_A$  behaupteten Zusammenhanges erklärt. Es wird dann die Wahrscheinlichkeit ermittelt, mit der diese Daten zufällig, d.h. also bei Zutreffen der Nullhypothese auftreten können. Unterschreitet diese Zufallswahrscheinlichkeit der Daten eine vorher festzulegende obere Grenze, das Verlässlichkeits- oder Signifikanzniveau  $\alpha$ , so wird die Nullhypothese verworfen, die Daten also als mit hinreichender Wahrscheinlichkeit nicht mehr zufällig angesehen und die Alternative (Forschungshypothese) „auf dem Verlässlichkeitsniveau  $\alpha$ “ angenommen, d.h. also vorbehaltlich einer Irrtumswahrscheinlichkeit von höchstens  $\alpha$ . Man sagt auch, das Ergebnis sei „auf dem Verlässlichkeitsniveau  $\alpha$  signifikant“.

Bei diesem Verfahren gibt es zwei Möglichkeiten, Fehler zu begehen:

**Fehler I. Art:** Man verwirft eine Nullhypothese, die an sich richtig ist. Dies tritt auf, wenn zufällig in der beobachteten Stichprobe solche Daten auftreten, die die Forschungshypothese (Alternative) scheinbar bestätigen. Die Wahrscheinlichkeit dieses Fehlers ist nach oben durch das Verlässlichkeitsniveau  $\alpha$  begrenzt.

**Fehler II. Art:** Man behält eine Nullhypothese bei, obwohl an sich die Alternative (Forschungshypothese) richtig ist. Ein tatsächlich bestehender Zusammenhang wird also nicht „entdeckt“, bzw. läßt sich nicht statistisch sichern. Die Wahrscheinlichkeit dieses Fehlers wird mit  $\beta$  bezeichnet; ihre Höhe ist meistens nicht zu bestimmen.

Das folgende Schema soll diese Zusammenhänge noch einmal verdeutlichen:

		Tatsächlicher Zustand der Natur: richtig ist die
Aufgrund der beobachteten Daten Nullhypothese entscheidet man sich für die	Nullhypothese	Alternative
	richtige Entscheidung zugunsten der Nullhypothese	Fehler II. Art, Wahrscheinlichkeit $\beta$
Alternative	Fehler I. Art, Wahrscheinlichkeit $\alpha$	richtige Entscheidung zugunsten der Alternative

Gegen den oben skizzierten „klassischen“ Neyman-Pearsonschen Signifikanztest als Entscheidungsverfahren für die Adoption neuer Erkenntnisse in einer empirischen Wissenschaft läßt sich eine ganze Reihe von Kritikpunkten vorbringen, von denen einige hier zunächst nur aufgezählt werden sollen:

1. Das Verfahren betrachtet nur die Wahrscheinlichkeit der Daten unter Annahme der Zufalls-(=Null-)Hypothese. Die Wahrscheinlichkeit der Daten unter Annahme der Alternative bleibt in den meisten Fällen der Praxis unbekannt und unberücksichtigt, desgleichen die Wahrscheinlichkeit der Hypothesen selbst, die den Forscher eigentlich mehr interessieren sollten.
2. Die a-priori-Wahrscheinlichkeiten der Hypothesen bleiben unberücksichtigt, d.h. also die Tatsache, daß dem Forscher schon vor Durchführung seiner Untersuchung oder Beobachtung möglicherweise die eine oder andere der beiden konkurrierenden Hypothesen wahrscheinlicher erschien (und wieviel wahrscheinlicher) als die andere. Dieser a-priori-Wahrscheinlichkeit wird in der Forschungspraxis zu wenig Beachtung geschenkt; unter dem Gesichtspunkt einer Forschungsökonomie sollte jede Untersuchung so geplant werden, daß die konkurrierenden Hypothesen zumindest subjektiv a priori gleich wahrscheinlich erscheinen. Statt dessen wird aber meistens so geplant, daß die Nullhypothese kaum eine Chance hat, da die Alternative ( $\Theta \neq \Theta_0$ ) alles umfaßt, außer eben  $\Theta_0$  selbst.

Nach den Gesetzen der Informationstheorie ist der erwartete Informationsgewinn am größten, wenn alle Alternativen gleich wahrscheinlich sind. Tatsächlich werden aber in der Praxis häufig Nullhypothesen aufgestellt, die von vornherein so formuliert sind, daß sie möglichst leicht zu verwerfen sind (sog. „Elefant-im-Porzellanladen“-Versuchsplan). Diese Technik, noch dazu verbunden mit der Gefahr einer Voreingenommenheit zugunsten der Alternative bei der Beobachtung und Auswertung der Daten, läßt die Anwendung des klassischen Signifikanztests sehr fragwürdig erscheinen.

Andererseits müssen wir - besonders bei historischer Betrachtung der Dinge - berücksichtigen, daß die klassischen Signifikanztests auf einen frequentistischen Wahrscheinlichkeitsbegriff aufbauen und daher so etwas wie eine a-priori-Wahrscheinlichkeit der Hypothese gar nicht verwenden können, wenn über diese Hypothese noch keine Daten vorliegen. Mit einem subjektiven (oder Propensitäts-)Wahrscheinlichkeitsbegriff ist das eher möglich.

3. In dem oben dargestellten Verfahren des klassischen Signifikanztests fehlt jede Art der Bewertung der möglichen vier Ausgänge der Entscheidung (s. obiges Schema). Selbstverständlich werden dem Forscher die beiden Arten „richtiger“ Entscheidungen in den meisten Fällen wünschenswerter erscheinen als die Fehler I. und II. Art, aber sicherlich gibt es unter diesen

noch Unterschiede in der Bewertung, die bei dem Entscheidungsverfahren völlig außer acht bleiben. Beispielsweise wird dem Forscher meistens die richtige Entscheidung zugunsten der Alternative (seiner ursprünglichen Forschungshypothese!) lieber sein als die zugunsten der Nullhypothese, und zwischen den beiden Fehlerarten hat er möglicherweise auch Präferenzen. All dies bleibt bei dem obigen Entscheidungsverfahren unberücksichtigt.

Wir wollen im folgenden einige alternative Entscheidungsverfahren kennenlernen, die versuchen, einige der Nachteile des klassischen Signifikanztests zu vermeiden.

### 3. *Sequentielle Testverfahren*

Legt man sich auf eine spezifische Alternativhypothese fest, so kann man auch das Risiko eines Fehlers II. Art kalkulieren, wie wir oben gesehen haben. Es ist dann naheliegend, die obere Grenze der Wahrscheinlichkeit seines Auftretens mit  $\beta$  genau so festzulegen wie die des Fehlers I. Art mit  $\alpha$ . Damit wird dann die Stichprobengröße  $n$ , die beim klassischen Signifikanztest vorher festgelegt war, zur Zufallsvariablen: Man beobachtet solange, bis eine der kritischen Grenzen  $a$  (zur Ablehnung der Nullhypothese  $H_0$ ) oder  $\beta$  (zur Ablehnung der Alternative  $H_A$ ) erreicht ist; je nach Ausfall der Stichprobe kann das einmal früher, einmal später der Fall sein. Es ist dies das Prinzip des Waldschen Sequentialtests: Nach jeder Einzelbeobachtung wird entschieden, ob  $H_0$  oder  $H_A$  verworfen oder ob weiter beobachtet werden soll. Wald (1947) hat ausgerechnet, daß bei diesem Verfahren bis zu 50% der Stichprobengröße gespart werden können, die zur Erreichung der gleichen Effektivität (d.h. zu Entscheidungen zwischen  $H_0$  und  $H_A$  bei gleichem  $\alpha$  und  $\beta$ ) mit klassischen Signifikanztests benötigt würden. Bei der Konstruktion der sequentiellen Testverfahren spielten also ökonomische Gesichtspunkte eine Rolle, was bei den klassischen Neyman-Pearson'schen Tests in weit geringerem Maße, (wenn überhaupt) der Fall war. (Die zuständigen US-Behörden betrachteten die Sequenztests dementsprechend auch als einen so bedeutsamen Fortschritt gegenüber den herkömmlichen, daß Wald sie erst nach dem Kriege veröffentlichen durfte.)

Bei den sequentiellen Testverfahren wird nicht nur das  $\alpha$ -, sondern auch das  $\beta$ -Risiko vorher festgelegt - nach dem bei der Besprechung der klassischen Signifikanztests Gesagten ist dies jedoch nur möglich, wenn eine präzise Alternativhypothese vorliegt, d.h. eine solche, die - wie die Nullhypothese - einen bestimmten Wert als möglichen Parameter festlegt und nicht - wie meistens üblich - einen ganzen Bereich.

Im echten Sequenztest kann aber eine Entscheidung zugunsten von  $H_0$  oder  $H_1$  mit dem gleichen Risiko  $\alpha$  oder  $\beta$  oft schon vor Erreichen dieser Stichprobengröße  $N$  gefällt werden, die im Falle des klassischen Tests erforderlich gewesen wäre, je nach Ausfall der beobachteten Stichprobe: Wenn beispielsweise in der Grundgesamtheit  $\Theta = \Theta_1$  ist, also  $H_1$  gilt, so könnten wir ja das Glück haben, eine Stichprobe zu beobachten, deren Kennwert  $M$  weit rechts von  $\Theta_1$  fällt, also erst recht weit rechts von  $a$ , so daß bereits bei viel kleineren Stichprobengrößen  $N$  (und damit größeren Streuungen der Verteilung der Stichprobenmittelwerte,  $\sigma / \sqrt{N}$ ) eine Entscheidung zugunsten von  $H_1$  mit einer Irrtumswahrscheinlichkeit kleiner als  $\alpha$  möglich ist. Ebenso könnten bei Zutreffen von  $H_1$  ( $\Theta = 0$ .) Stichproben mit Mittelwerten  $M$  unterhalb von  $d$  auftreten, für die Entsprechendes bei der Entscheidung zugunsten von  $H_0$  gilt.

Die Strategie der Sequenztests ist nun - anschaulich gesprochen - bei der Situation in Abb. 2a zu beginnen und nach jedem beobachteten Stichprobenelement zu entscheiden, ob bereits entweder

- (a) eine Entscheidung zugunsten von  $H_1$  getroffen werden kann mit einer Irrtumswahrscheinlichkeit kleiner als  $\alpha$ , weil der Mittelwert  $M$  der bisher beobachteten Stichprobe rechts von  $d_0$  liegt, oder ob
- (b) eine Entscheidung zugunsten von  $H_0$  getroffen werden kann mit einer Irrtumswahrscheinlichkeit kleiner als  $\beta$ , weil der Mittelwert  $M$  der Stichprobe links von  $d_1$  liegt, oder ob
- (c) der Mittelwert  $M$  der bisher beobachteten Stichprobe zwischen  $d_1$  und  $d_0$  liegt, damit eine Entscheidung mit Irrtumswahrscheinlichkeiten entsprechend den vorgegebenen Werten  $\alpha$  und  $\beta$  (oder kleiner) noch nicht möglich ist, und daher die Stichprobe weiter vergrößert werden muß.

Statt in zwei Teilräumen wie bei der klassischen Statistik ( $H_0$  annehmen unterhalb von  $d$ ,  $H_1$  oberhalb von  $d$ ) unterteilen wir den Bereich möglicher Stichproben(kennwerte) hier in drei:  $H_0$  annehmen unterhalb von  $d_1$ ,  $H_1$  annehmen oberhalb von  $d_0$ , weiterprüfen zwischen  $d_0$  und  $d_1$ . Um zu dieser Entscheidung zwischen (a), (b) und (c) zu kommen, betrachten wir den Quotienten aus der Wahrscheinlichkeit der Daten (Stichprobe) bei Zutreffen von  $H_1$ ,  $P(D | H_1)$ , durch die Wahrscheinlichkeit der Daten bei Zutreffen von  $H_0$ ,  $P(D | H_0)$ , den sog. Likelihood-Quotienten  $L$ , der schon vor der Erfindung der Sequentialtests von Fisher (1921) als Test zur Entscheidung zwischen zwei konkurrierenden Hypothesen eingeführt wurde:

$$L(D) = \frac{P(D | H_1)}{P(D | H_0)} = \frac{L(H_1 | D)}{L(H_0 | D)}$$

Der darin verwendete (von Fisher eingeführte) Begriff der Likelihood (s. z.B. Fisher, 1959, S. 68ff.) ist dabei als Funktion von zwei Variablen abhängig, nämlich von den Daten  $D$  und von den Hypothesen  $H_i$ , genauer: von den hypothetischen Parametern. Fisher selbst benutzte den Likelihood-Begriff als

Übergang vom Bereich der Ereignisse (Daten) in den Bereich der Hypothesen: Die Wahrscheinlichkeit  $P(D \mid H_i)$  der gegebenen Daten  $D$  unter der Annahme, daß die Hypothese  $H_i$  richtig ist, kann man als Funktion der Daten  $D$  betrachten. Man kann aber auch die Daten als fest vorgegeben ansehen und ihre Wahrscheinlichkeit  $P(D \mid H_i)$  als Funktion der Hypothesen  $H_i$ , beispielsweise als Funktion eines variierenden hypothetischen Parameters  $\Theta_i$ . Genau das letztere geschieht im Likelihood-Konzept: Man betrachtet die Wahrscheinlichkeit der Daten,  $P(D \mid H_i)$ , als Funktion der möglichen Hypothesen  $H_i$  (s. z.B. Box & Tiao, 1973, S. 10ff.). Diese Likelihood  $L(H_i \mid D)$  darf aber nicht mit der Wahrscheinlichkeit der Hypothesen selbst  $P(H_i)$  oder  $P(H_i \mid D)$  verwechselt werden. In der (später näher zu besprechenden) Bayes-Statistik dient die Likelihood dann als Werkzeug zur Berechnung der Wahrscheinlichkeiten  $P(H \mid D)$  der Hypothesen a posteriori (d.h. nach Kenntnis der Daten) aus den Wahrscheinlichkeiten  $P(H)$  der Hypothesen a priori (d.h. vor Kenntnis der Daten), also zu einer Transformation im Bereich der Hypothesen.

Wie wir sehen werden, kann aufgrund des Likelihood-Quotienten  $L(D)$  entschieden werden, in welchem der drei Teilräume (a), (b) oder (c) wir uns mit einer Stichprobe  $D$  befinden. Erreicht oder überschreitet  $L(D)$  einen gewissen Wert  $A$ , ist also die Wahrscheinlichkeit der Daten unter Annahme von  $H_1$  (Zähler von  $L(D)$ ) im Verhältnis zu ihrer Wahrscheinlichkeit unter Annahme von  $H_0$  (Nenner von  $L(D)$ ) größer oder gleich  $A$ , so nehmen wir  $H_1$  an. Erreicht oder unterschreitet  $L$  andererseits einen gewissen Wert  $B$ , so nehmen wir  $H_0$  an. Liegt  $L(D)$  zwischen  $A$  und  $B$ , so befinden wir uns im Teilraum zwischen  $d_1$  und  $d_0$ , müssen also noch weitere Daten erheben.

Es kommt nun darauf an, passende Werte für  $A$  und  $B$  zu finden, bei denen die oben genannten Entscheidungen genau mit den gewählten Risiken  $\alpha$  und  $\beta$  gefällt werden können.

Der Wert  $A$  soll von dem Likelihood-Quotienten  $L(D)$  genau dann überschritten werden, wenn zu der bisherigen Stichprobe eine Beobachtung hinzukommt, die zur Verwerfung von  $H_0$  mit einer Irrtumswahrscheinlichkeit  $\leq \alpha$  führt. Gleichzeitig soll die Wahrscheinlichkeit des Auftretens dieser Stichprobe unter Annahme von  $H_1$ , also  $P(D \mid H_1)$ , kleiner oder gleich  $(1 - \beta)$  sein. Also gilt für den Wert  $A$ :

$$A = \frac{1 - \beta}{\alpha}$$

Durch analoge Überlegungen finden wir, daß für  $B$  gelten muß:

$$B = \frac{\beta}{1 - \alpha}$$

Die Entscheidung für den Sequenztest bei vorgegebenem  $\alpha$  und  $\beta$  lautet also:

Wähle  $H_1$ , falls

$$L(D) = \frac{P(D | H_1)}{P(D | H_0)} \geq A = \frac{1-\beta}{\alpha},$$

wähle  $H_0$ , falls

$$L(D) = \frac{P(D | H_1)}{P(D | H_0)} \leq B = \frac{\beta}{1-\alpha},$$

teste weiter, solange

$$A > L(D) > B \text{ ist.}$$

Zu diesem Verfahren noch einige technische Anmerkungen:

- (1) Sollte man eine Stichprobe bekommen, mit der man zu lange (d.h. bis zu einer größeren Anzahl Beobachtungen  $N$ , als ursprünglich maximal vorgesehen) mit dem Likelihoodquotienten  $L(D)$  in dem Raum zwischen  $A$  und  $B$  bleibt, so kann der Test vorzeitig abgebrochen werden. In Wald (1947) ist für diesen Fall angegeben, wie man ausrechnen kann, um wieviel sich die verbleibende Irrtumswahrscheinlichkeit erhöht, falls man auf dieser Basis eine Entscheidung fällt.
- (2) Der Likelihoodquotient  $L(D)$  ist im Falle diskreter Daten, z.B. binomialverteilten Beobachtungen, der Quotient aus den Wahrscheinlichkeiten der Daten unter den beiden konkurrierenden Hypothesen, bei Daten auf einem Kontinuum (z.B. normalverteilten Daten) kann  $L(D)$  auch der Quotient der entsprechenden Wahrscheinlichkeitsdichten sein.
- (3) Sind die einzelnen Beobachtungen  $X_1, X_2, \dots, X_N$  der Stichprobe  $D = \{X_1, X_2, \dots, X_N\}$  voneinander unabhängig, wie es meistens angenommen werden kann, so ist die Wahrscheinlichkeit der Stichprobe  $D$  gleich dem Produkt der Wahrscheinlichkeiten der einzelnen Beobachtungen. Aus praktischen Gründen empfiehlt es sich dann, die Ungleichungen der obigen Entscheidungsregel zu logarithmieren. Zu dem Logarithmus des Likelihoodquotienten,  $\log L(\{X_1, \dots, X_{N-1}\})$  der bisherigen Stichprobe wird dann der Logarithmus der neuen ( $N$ -ten) Beobachtung  $X_N$ ,  $\log L(X_N) = \log P(X_N | \Theta_1) - \log P(X_N | \Theta_0)$  hinzuaddiert und geprüft, ob die neue Summe die Grenze  $\log A = \log(1-\beta) - \log \alpha$  überschreitet oder die Grenze  $\log B = \log \beta - \log(1-\alpha)$  unterschreitet.

Die Entwicklung der theoretischen Grundlagen des Sequenztests sind z. B. bei Wald (1947, 1950) oder Fisz (1970) nachzulesen, eine Reihe Hinweise zur praktischen Durchführung gibt u.a. Moroney (1956), darunter z.B. auch auf Möglichkeiten einer graphischen Darstellung des Tests, bei der in einem Koordinatensystem zwischen der Stichprobengröße  $N$  (Abszisse) und einer Funktion der Stichprobenbeschaffenheit, z.B. der Summe der bisherigen Meßwerte



(Ordinate) verfolgt werden kann, ob und wann der die Stichprobe repräsentierende Punkt eine der eingezeichneten Grenzen überschreitet, jenseits deren zugunsten einer der Hypothesen  $H_0$  oder  $H_1$  entschieden werden kann. Im deutschen Sprachraum findet man Darstellungen von Sequenztests bei Weber (1972) und bei Lienert (1978).

Wir sind auf die Sequenztests in diesem Rahmen etwas näher eingegangen, zum einen, wie schon gesagt, um ihre Anwendung zu fördern, zum anderen, weil sie einen bedeutenden Schritt von den klassischen Signifikanztests zu den nun zu besprechenden neueren Entscheidungsverfahren darstellen, in denen mehr Aspekte der Entscheidungssituation als die Risiken  $\alpha$  und  $\beta$  berücksichtigt werden sollen.

Auch die Sequenztests lassen die Frage offen, woher man denn nun eigentlich die vorher festzulegenden Grenzen der Irrtumswahrscheinlichkeit  $\alpha$  und  $\beta$  nehmen soll, oder wie bestimmte Werte von  $\alpha$  und  $\beta$  zu rechtfertigen sein könnten. Eine Möglichkeit, dieser Frage zu entgehen, die man auch in der neueren Literatur gelegentlich genutzt findet, besteht darin, gar keine kritischen Grenzen anzugeben (und damit auch keine Entscheidung zu fallen), sondern nur deskriptiv für jede der konkurrierenden Hypothesen die Wahrscheinlichkeit  $P(D | H_i)$  der beobachteten Daten im Falle der Gültigkeit der betreffenden Hypothese  $H_i$  anzugeben, oder besser die Wahrscheinlichkeit  $P(H_i | D)$  und dem Leser dann das Urteil selbst zu überlassen.

In vielen Fällen ist das sicherlich informativer als die lapidare Feststellung, ein Ereignis sei „signifikant auf dem . . . ( $\alpha$ )-Verlässlichkeitsniveau“.

Allerdings läßt sich eine solche Hypothesen-Wahrscheinlichkeit auf Basis des bisher verwendeten frequentistischen (klassischen) Wahrscheinlichkeitsbegriffs kaum konzipieren, es sei denn über die Vorstellung einer Vielzahl gleichgelagerter Fälle, in denen induktiv man von Stichprobendaten auf die relativen Häufigkeiten der zugrundeliegenden bestimmten Parameterwerte (= Hypothesen) schließt. Hier läßt sich ein subjektiver oder der Poppersche Propensitäts-Wahrscheinlichkeitsbegriff sinnvoller einsetzen.

Hat man schon vor der Untersuchung irgendwelche Vorstellungen von der Zutreffenswahrscheinlichkeit  $P(H_i)$  der einzelnen konkurrierenden Hypothesen gehabt, so können die (als subjektive interpretierten) Wahrscheinlichkeiten  $P(H_i | D)$  nach dem Satz von Bayes ausgerechnet werden:

$$P(H_i | D) = \frac{P(D | H_i) \cdot P(H_i)}{\sum_i P(D | H_i) P(H_i)},$$

wobei die Summierung im Nenner über alle in Frage kommenden konkurrierenden (d.h. einander ausschließenden und gemeinsam alle Möglichkeiten erschöpfenden) Hypothesen geht.

Ein anderer Weg besteht darin, daß man versucht, rationale Kriterien für die Bewertung der verschiedenen möglichen Ausgänge der Entscheidung zu finden, d.h. eine Zuordnung von Nutzwerten oder Kosten (Schadenswerten oder negativen Nutzwerten) zu den einzelnen Konsequenzen der Entscheidung, z.B. für die Folgen der Fehlentscheidung, aufgrund einer Stichprobe  $H_0$  anzunehmen, während in der Grundgesamtheit  $H_1$  richtig ist, usw.

Wir werden auf diese Möglichkeiten im Rahmen der Darstellung der Bayes-Statistik noch näher eingehen.

#### 4. *Likelihood-Quotienten-Test*

Mit den sequentiellen Testverfahren nahe verwandt (und eigentlich ein Vorläufer von ihnen) ist der Likelihood-Quotienten-Test, bei dem der aus dem vorigen Abschnitt bekannte Likelihood-Quotient

$$L(D) = P(D \mid H_1) / P(D \mid H_0) = \frac{L(H_1 \mid D)}{L(H_0 \mid D)}$$

mit einem vorher festgelegten Kriterium verglichen wird, um über die Annahme einer der beiden konkurrierenden Hypothesen zu entscheiden (Stegmüller, 1973, S. 167). Im Unterschied zu den sequentiellen Verfahren wird diese Prüfung aber erst nach vollständiger Stichprobenerhebung (und nicht nach jedem Einzeldatum) erhoben, und das vorher festzulegende Kriterium kann, muß aber nicht von den zulässigen Irrtumswahrscheinlichkeiten  $\alpha$  und  $\beta$  abhängen. Statt dessen können zur Festsetzung des Kriteriums beispielsweise auch ökonomische Gesichtspunkte herangezogen werden, wie etwa die Kosten und Nutzen der Konsequenzen der anstehenden Entscheidung.

Die Berücksichtigung solcher Überlegungen sowie auch die der sog. a-priori-Wahrscheinlichkeiten der Hypothesen  $P(H_0)$  und  $P(H_1)$  - diese dabei meistens in subjektiver Interpretation - ist Gegenstand der im folgenden zu besprechenden Bayes-Verfahren. Bei entsprechender Wahl des Kriteriums kann deshalb jeder Likelihood-Quotienten-Test auch als Bayes-Strategie betrachtet werden (s. auch Bredenkamp, 1972, S. 139; sowie Chernoff & Moses, 1959, S. 248).

Auf einen Unterschied zwischen den Waldschen Sequenztests und den Likelihood-Tests soll noch hingewiesen werden: Im ersteren werden jeweils die Wahrscheinlichkeiten der beobachteten Daten und aller extremeren (noch stärker die jeweilige Hypothese begünstigenden) Daten betrachtet, damit diese mit den als Kriterium herangezogenen Verteilungsschwänzen  $\alpha$  und  $\beta$  verglichen werden können; im letzteren wird nur jeweils die Wahrscheinlichkeit bzw. Wahrscheinlichkeitsdichte der beobachteten Daten selbst berücksichtigt.

## 5. Bayes-Statistik

Wie wir gesehen haben, konzentriert sich die Kritik an der Anwendung der klassischen Neyman-Pearsonschen Testtheorie hauptsächlich auf diese Punkte:

1. Das Fehlen einer spezifischen, brauchbaren Alternativhypothese in den meisten Fällen, und damit verbunden
2. die Unmöglichkeit, über das Beta-Risiko etwas auszusagen,
3. die unangemessene Verquickung von Signifikanzhöhe und Effektstärke, und damit zusammenhängend
4. der Mangel an Kriterien für die Wahl eines Verlässlichkeitsniveaus, und
5. die Unmöglichkeit, etwas darüber auszusagen, wie brauchbar die aufgrund der Daten besser gestützte Hypothese wirklich ist.

Einige dieser Mängel - insbesondere die erstgenannten - versucht Witte (1980) in einer Erweiterung des klassischen Modells durch Hinzunahme eines Ansatzes von Cohen (1977) zur Entflechtung von Effektstärke und Verlässlichkeitsniveau zu beheben: Die Stichprobengröße wird dabei in Abhängigkeit von gesuchter Effektstärke und gewünschtem Alpha- und Beta-Risiko gewählt. Dieser Ansatz ermöglicht auch die Integration der Ergebnisse mehrerer Einzeluntersuchungen (Replikationen) zu einer Gesamtentscheidung über konkurrierende Hypothesen, gibt aber dem Forscher immer noch keine Hilfe bei der Wahl des Alpha- und Beta-Risikos. Hierin unterscheidet sich von den bisher besprochenen Ansätzen der von Bayes (1763, 1958), der auf die Festlegung solcher Risikoschranken ganz verzichtet und die Entscheidung zwischen konkurrierenden Hypothesen ganz an ihrem Erwartungswert orientiert.

### 5.1 Vorgehensweise der Bayes-Statistik

Einer der wichtigsten Unterschiede der Bayes'schen statistischen Verfahren gegenüber den klassischen besteht darin, daß hier zur Bewertung der konkurrierenden Hypothesen (oder Entscheidungen zwischen ihnen) auch Informationen mitverwendet werden, die nicht auf der beobachteten Stichprobe selbst beruhen, sondern schon vorher verfügbar waren. Sie gehen als a-priori-Wahrscheinlichkeiten in die Berechnung der Wahrscheinlichkeiten der Hypothesen nach Beobachtung der Stichprobe ein, welche im wesentlichen dem Satz von Bayes folgt:

Die Wahrscheinlichkeit  $P(H_i | D)$  einer Hypothese  $H_i$  nach Kenntnis der Daten  $D$  ist gleich ihrer Wahrscheinlichkeit  $P(H_i)$  ohne Kenntnis dieser Daten, multipliziert mit der Wahrscheinlichkeit  $P(D | H_i)$  dieser Daten  $D$  unter der Voraussetzung, daß die Hypothese  $H_i$  zutrifft, geteilt durch die totale Wahrscheinlichkeit  $P(D)$ , daß diese Daten überhaupt auftreten:

$$P(H_i | D) = P(H_i) \cdot P(D | H_i) / P(D),$$

wobei

$$P(D) = \sum P(D | H_j) P(H_j)$$

die Summe der Wahrscheinlichkeiten  $P(D | H_j)$  der Daten unter allen möglichen konkurrierenden, d.h. einander ausschließenden und alle Möglichkeiten erschöpfenden, Hypothesen ist, jeweils gewichtet mit den a-priori-Wahrscheinlichkeiten der betreffenden Hypothesen.

Der Beweis des Satzes von Bayes ergibt sich unmittelbar aus der Definition der bedingten Wahrscheinlichkeit:

$$(1) \quad P(D | H) = P(D \& H) / P(H)$$

$$(2) \quad P(H | D) = P(D \& H) / P(D)$$

Aus (1) folgt:

$$(3) \quad P(D \& H) = P(D | H) P(H)$$

Dies eingesetzt in (2) ergibt:

$$(4) \quad P(H | D) = P(D | H) P(H) / P(D), \text{ q.e.d.}$$

## 5.2 Robustheit der Schätzung (principle of stable estimation)

Ein besonderes Problem bei der Anwendung Bayes'scher statistischer Verfahren entsteht aus der Notwendigkeit, für jede der konkurrierenden Hypothesen  $H_j$  eine a-priori-Wahrscheinlichkeit  $p(H_j)$  zu finden. In der Forschungspraxis gilt es, dazu das in fast allen Fällen mehr oder weniger latent vorhandene Vorwissen des Wissenschaftlers oder Experten (oder seiner Kollegen) zu erfassen und in eine Form zu bringen, in der es in die Bayes'sche Formel als Verteilung der  $P(H_j)$  eingesetzt werden kann - im Grunde handelt es sich um ein Skalierungsproblem.

Gerade an diesem Punkt hat heftige Kritik an den Bayes-Verfahren angesetzt. Ihre Gegner befürchten, daß hier subjektive und Vor-Urteile in den Datenanalyse-Prozeß einfließen und das Ergebnis verfälschen. Dem wird von den Verfechtern der Bayes-Verfahren entgegengehalten, daß der Schaden so groß nicht werden könnte:

Auch erheblich von den „wahren“ Werten abweichende a-priori-Wahrscheinlichkeiten  $P(H_j)$  für die Hypothesen  $H_j$  werden durch auch nur einigermaßen aussagekräftige Daten sehr rasch zu erheblich „richtigeren“ Hypothesenwahr-

scheinlichkeiten  $P(H \mid D)$  korrigiert. Dies ist das „principle of stable estimation“, das an dem folgenden Beispiel demonstriert werden soll:

Ein Hochschullehrer habe die Hypothese  $H_1$ , daß das Seminar eines seiner Kollegen unverhältnismäßig stark von Damen frequentiert wird, jedenfalls meint er, daß 80% der Hörer dieses Seminars weiblichen Geschlechts seien, während in der Grundgesamtheit nur 50% der Studierenden des betreffenden Fachs weiblichen Geschlechts sind. Besagter Hochschullehrer ist sehr stark von der Richtigkeit seiner Hypothese überzeugt; er hält sie mit  $P(H_1 : \Pi = 0.8) = 0.9$  für neunmal so wahrscheinlich wie  $P(H_0 : \Pi = 0.5) = 0.1$ . Wir wollen dieses Problem einmal sequentiell angehen (obwohl das kein typisches Merkmal der Bayes-Statistik ist) und nach jedem neu beobachteten Datum (= Geschlecht eines Seminarteilnehmers) die Hypothesenwahrscheinlichkeiten  $P(H_1 \mid D)$  und  $P(H_0 \mid D)$  oder einfacher ihr Verhältnis  $P(H_1 \mid D) / P(H_0 \mid D)$  neu betrachten (weil sich in diesem Falle  $P(D)$  wegekürzt), wobei wir annehmen, daß die Reihenfolge des Eintreffens im Seminarraum ein echter Zufallsprozeß mit gleichen Bedingungen für beide Geschlechter sei:

In der folgenden Tabelle 1 gibt die erste Spalte jeweils das Geschlecht des hereinkommenden Seminarteilnehmers (m oder w) an, die zweite Spalte den Likelihoodquotienten dieser Beobachtung:  $P(D \mid H_1) / P(D \mid H_0)$ , mit dem das Verhältnis der Hypothesenwahrscheinlichkeit der vorangehenden Zeile (vor Beobachtung dieses Datums) multipliziert wird. Dabei ist  $P(D \mid H_0)$  in jedem Falle 0.5,  $P(m \mid H_1) = 0.2$ ,  $P(w \mid H_1) = 0.8$ . Die dritte Spalte gibt das aufgrund des Datums veränderte (revidierte) Verhältnis der Hypothesenwahrscheinlichkeiten  $P(H_1 \mid D) / P(H_0 \mid D)$  an.

Tabelle 1:

Lfd. Nr.	Datum D: m oder w	$\frac{P(D \mid H_1)}{P(D \mid H_0)}$	$\frac{P(H_1 \mid D)}{P(H_0 \mid D)}$
0	-	-	9 (a-priori-Verhältnis)
1	w	1.6	14.4
2	m	0.4	5.76
3	m	0.4	2.304
4	w	1.6	3.6864
5	w	1.6	5.89824
6	m	0.4	2.359296
7	m	0.4	0.9437184
8	m	0.4	0.3774873
9	w	1.6	0.6039796
10	w	1.6	0.9663673
11	m	0.4	0.3865469
12	w	1.6	0.6184750

Die Tabelle zeigt deutlich, wie stark das Verhältnis der Hypothesenwahrscheinlichkeiten auf die Revision durch die Daten reagiert; trotz der stark einseitigen a-priori-Festlegung mit 9:1 zugunsten von  $H_1$  hat sich das Verhältnis schon nach 7 Beobachtungen umgekehrt (d.h. 1 unterschritten, damit höhere Wahrscheinlichkeit für  $H_0$ ).

Ein drastischeres Beispiel der Kraft des „principle of stable estimation“ (das hier darzustellen zuviel Platz nehmen würde) findet man in Hofstätter & Wendt, 1974, S. 268; die allgemeinen Bedingungen und Grundlagen dieses Prinzips sind bei Edwards, Lindman & Savage (1963) ausführlicher dargestellt, die mathematischen Grundlagen bei Blackwell & Dubins (1962).

Wichtige Voraussetzungen für seine Gültigkeit sind, daß die Daten hinreichend aussagekräftig sind, und daß die a-priori-Wahrscheinlichkeitsverteilung über die Daten einigermaßen gleichmäßig ist, und nicht etwa gerade der „richtigen“ Hypothese eine extrem niedrige und einer falschen Hypothese eine übermäßig hohe a-priori-Wahrscheinlichkeit gibt. Die Aussagekraft der Daten hängt mit dem Inhalt der Hypothesen zusammen: Die Beobachtung, daß ein Teilnehmer des Seminars männlichen Geschlechts ist, hat bei dem Hypothesenpaar  $H_1 : \Pi = 0.6 / H_0 : \Pi = 0.5$  weniger Aussagekraft als bei dem Hypothesenpaar  $H_2 : \Pi = 0.2 / H_3 : \Pi = 0.8$ , wie man mit Hilfe der entsprechenden Rechnungen leicht selbst demonstrieren kann. Wir kommen in dem allgemeineren Fall der Parameterschätzung mit Bayes-Verfahren noch einmal darauf zurück.

### 5.3 Vergleich mit der klassischen Statistik

Gerade wegen der oben dargestellten starken Reaktion der Hypothesenwahrscheinlichkeit auf Daten bei der Revision nach dem Satz von Bayes ist dem Verfahren von seinen Gegnern häufig der Vorwurf gemacht worden, es sei allzu empfindlich und würde „radikaler“ arbeiten als die klassischen Testverfahren, also eher zur Verwerfung der Nullhypothese und Annahme der Alternativhypothese führen als diese. Edwards, Lindman & Savage (1963) demonstrierten an einem Beispiel, daß das nicht der Fall sein muß, sondern eher das Gegenteil eintreten kann (S. 221 f.): Nehmen wir den Fall eines zweiseitigen t-Tests mit hinreichend vielen Freiheitsgraden, so daß wir von der Normalverteilungs-Approximation der t-Verteilung Gebrauch machen können. Unter Annahme der Nullhypothese erwarten wir, daß die Daten mit einer Wahrscheinlichkeit von 0.025 t-Werte über 1.96 liefern werden, und mit einer Wahrscheinlichkeit von 0.005 t-Werte über 2.58. (Analoges gilt für t-Werte unter -1.96 und unter -2.58 am anderen Ende der Verteilung.) Wenn die Nullhypothese zutrifft, liegt der t-Wert also mit einer Wahrscheinlichkeit von  $0.025 - 0.005 = 0.02$  zwischen 1.96 und 2.58.

Nehmen wir an, der (exakten) Nullhypothese stehe eine diffuse Alternativhypothese gegenüber, bei der alle t-Werte von -20 bis +20 gleichwahrscheinlich sind, mit Ausnahme des Wertes 0. Wenn diese Alternativhypothese zutrifft, liegt der t-Wert mit einer Wahrscheinlichkeit von  $(2.58 - 1.96) / (+20 - (-20)) = 0.0155$  in dem oben betrachteten Intervall von 1.96 bis 2.58. Das bedeutet, daß die Wahrscheinlichkeit der Daten (aus denen ja der t-Wert berechnet wurde, die er also repräsentiert) unter Annahme der Alternativhypothese mit 0.0155 weniger wahrscheinlich ist als unter Annahme der Nullhypothese mit 0.02 - obwohl diese Nullhypothese bei Anwendung der klassischen Testverfahren auf dem 0.05-Verlässlichkeitsniveau verworfen werden würde. Das Bayes-Verfahren dagegen trägt dieser geringeren Datenwahrscheinlichkeit unter der Alternativhypothese Rechnung.

Diesem vielzitierten Beispiel ist allerdings von Gegnern der Bayes-Statistik der Vorwurf gemacht worden, daß es sehr künstlich konstruiert sei; auf dieser Basis ließe sich zu jeder verworfenen Nullhypothese eine noch unwahrscheinlichere Alternativhypothese finden (Witte, 1980).

## 5.4 Integration von Daten aus verschiedenen Quellen

Ein zusätzlicher Vorteil der Bayes-Revision von Wahrscheinlichkeiten besteht darin, daß mit ihrer Hilfe relativ einfach Daten aus den verschiedensten Quellen integriert und gemeinsam zur Revision der Hypothesen-Wahrscheinlichkeiten herangezogen werden können. Hat man beispielsweise in einer Untersuchung der Effektivität von Lehrplänen eine Reihe abhängiger Variablen erhoben, wie z.B. Durchschnittsnoten in bestimmten Unterrichtsfächern, durchschnittlich erzielte Testwerte in verschiedenen Leistungstests sowie die Anzahl der Versager oder Sitzenbleiber unter den konkurrierenden Lehrplänen, so ist es im Rahmen der klassischen Statistik relativ kompliziert und mühsam, diese verschiedenen Datenquellen zu einer Gesamtbeurteilung eines etwaigen Unterschiedes zwischen den verschieden behandelten Gruppen heranzuziehen.

Bei der Anwendung eines Bayes-Verfahrens dagegen braucht man nur die Datenwahrscheinlichkeiten  $P(D | H_i)$  jeder Datenquelle unter den konkurrierenden Hypothesen zu kennen (die man bei den klassischen Verfahren, zumindest unter der Nullhypothese, ja auch benötigt), und kann diese ohne Rücksicht auf ihre Herkunft zur Revision von  $P(H_i)$  hintereinander multiplizieren, vorausgesetzt, sie sind unabhängig.

Dies gilt übrigens für Likelihood- und Sequenztest unter den entsprechenden Bedingungen in gleicher Weise bei entsprechender subjektivistischer oder Propensitäts-Interpretation der Wahrscheinlichkeiten.

## 6. Parameter-Schätzung

In den vorangegangenen Abschnitten haben wir es mit Entscheidungen zwischen diskreten Hypothesen zu tun; die Menge  $A$  der Alternativen  $H_0$ ,  $H_1$  usw., aus der eine ausgewählt werden sollte, war relativ klein; meistens enthielt sie nur zwei Elemente:  $A = \{H_0, H_1\}$ . Die gleichen Entscheidungsprinzipien lassen sich aber ohne weiteres auf den Fall großer Alternativenmengen verallgemeinern, im Extremfall auch auf Mengen mit unendlich vielen Alternativen, nämlich möglichen Parameterwerten auf einem Kontinuum. Wir haben es dann mit einem Fall von Parameterschätzungen zu tun: Im Prinzip handelt es sich dabei um Entscheidungen für die jeweils beste Alternative, d.h. den jeweils besten Wert aus dem Kontinuum möglicher Parameterwerte.

Auch bei den bisher besprochenen Verfahren zur Entscheidung zwischen konkurrierenden Hypothesen ließen sich die Hypothesen selbst in den meisten Fällen durch Parameterwerte identifizieren oder operationalisieren, z.B. durch  $H_0 : \Theta = \Theta_0$  und  $H_1 : \Theta = \Theta_1$ . Nehmen wir jetzt an, daß wir mehr als zwei solche konkurrierenden Hypothesen haben, etwa  $H_0 : \Theta = \Theta_0$ ,  $H_1 : \Theta = \Theta_1$ ,  $H_2 : \Theta = \Theta_2, \dots, H_n : \Theta = \Theta_n$ , und daß diese möglichen Parameterwerte  $\Theta_i$  alle auf einem Kontinuum liegen, wobei auch noch beliebige Zwischenwerte als hypothetische Parameter möglich sind, so ist aus der Entscheidung zwischen konkurrierenden Hypothesen eine Parameterschätzung geworden.

An die Stelle der bisher betrachteten diskreten Wahrscheinlichkeitsverteilungen  $P(H)$ ,  $P(H | D)$  über Hypothesen, bzw.  $P(D | H)$ ,  $P(D)$  über Daten treten jetzt kontinuierliche (möglichst stetige) Verteilungen, und als Werkzeug zur Auswahl des Optimums, d.h. des Parameterwertes (= der Hypothese) mit maximalem Nutzen (oder minimalem Schaden) tritt die Differentialrechnung hinzu.

Entsprechend den oben genannten verschiedenen Verfahren zur Entscheidung zwischen diskreten Hypothesen sind auch für den kontinuierlichen Fall - also die Parameterschätzung - unterschiedliche Verfahren entwickelt worden.

### 6.1 Lösung der kleinsten Quadrate

Der klassischen Statistik (im diskreten Fall) entspricht am ehesten das Prinzip der „kleinsten (Abweichungs-)Quadrate“. Hierbei wird derjenige Parameterwert gewählt, bei dem die Summe der Quadrate der Abweichungen der tatsächlich beobachteten Daten  $X_{ij}$  von den aufgrund der Parameter vorhergesagten Werten  $\xi_{ij}$  minimal ist. Ein sehr einfaches Beispiel für eine Parameterschätzung nach dem Prinzip der kleinsten Quadrate ist die Schätzung eines Populationsmittelwertes durch das arithmetische Mittel einer Stichprobe; ein anderes



Beispiel ist die Schätzung der Parameter einer Regressionslinie zwischen korrelierenden Daten. Praktisch wird bei solchen Parameterschätzungen die Summe der quadrierten Abweichungen,  $\sum (x_{ij} - \xi_{ij})^2$ , als Funktion des (der) Parameter(s)  $\Theta_i$  betrachtet:  $\xi_{ij} = f(\Theta_i)_j$ , und von dieser Summe  $\sum (x_{ij} - f(\Theta_i)_j)^2$  die

1. Ableitung nach dem Parameter  $\Theta_i$  gebildet, diese gleich Null gesetzt:

$$\frac{d (\sum (x_{ij} - f(\Theta_i)_j)^2)}{d \Theta_i} = 0,$$

und dann in der 2. Ableitung geprüft, ob ein Maximum oder Minimum vorliegt. Entsprechendes gilt bei mehreren Parametern; hier werden die partiellen Ableitungen nach den einzelnen Parametern gleich Null gesetzt. (Für nähere Einzelheiten siehe z.B. Hofstätter & Wendt (1974), Kap. 18 (S. 252 ff.).)

## 6.2 Maximum-Likelihood-Schätzung

Dem Likelihoodquotienten-Test entspricht am ehesten die Maximum-Likelihood-Schätzung. Bei diesem Prinzip wird wieder die Wahrscheinlichkeit bzw. Wahrscheinlichkeitsdichte der Daten in Abhängigkeit vom Parameterwert betrachtet, dieser dann als stetige Variable behandelt und die Wahrscheinlichkeitsdichte der gegebenen Daten in Abhängigkeit von den Parameterwerten als Funktion betrachtet, diese wiederum nach den Parametern differenziert, und die erste Ableitung nach jedem Parameter gleich Null gesetzt. Aus den so entstandenen Gleichungen werden die Parameterwerte berechnet. Sie stellen diejenigen Parameterwerte dar, bei denen die Wahrscheinlichkeit der beobachteten Daten maximal ist. Maximiert wird also  $P(x \mid \Theta)$  in Abhängigkeit von  $\Theta$  bei gegebenem  $x$ .

Mit der Bayes-Statistik lassen sich dagegen Parameterwerte schätzen, deren Wahrscheinlichkeit bei den beobachteten Daten maximal ist; maximiert wird also  $P(\Theta \mid x)$  bei gegebenem  $x$ . Außerdem geht hier zusätzlich noch die a-priori-Verteilung über den Parameter,  $P(\Theta)$ , mit ein. Die Verteilung a priori,  $P(\Theta)$ , wird unter Anwendung des Satzes von Bayes (bzw. seiner modifizierten Form für den Fall kontinuierlicher Verteilungen) revidiert aufgrund der Datenwahrscheinlichkeit(sichte)  $P(x \mid \Theta)$ , und dann der wahrscheinlichste Parameterwert gewählt. Das Verfahren ist im Grunde der Maximum-Likelihood-Schätzung sehr ähnlich, und das Ergebnis ihr insofern proportional, als hier ebenfalls  $P(x \mid \Theta)$  eingeht. Allerdings ist eine Verschiebung des Ergebnisses noch aufgrund der ebenfalls berücksichtigten Vor-Wahrscheinlichkeiten  $P(\Theta)$  möglich.

## 6.3 Konjugierte Verteilungen

Bei dieser Revision von a-priori-Verteilungen über den Raum möglicher Parameter durch die Verteilung (oder genauer: durch Kennwerte der Verteilung) der Daten in eine a-posteriori-Verteilung spielt es eine besonders wichtige Rolle, daß man sog. konjugierte Verteilungen findet, d.h. solche, die bei der Revision aus einer bekannten theoretischen Verteilung nach Möglichkeit in die gleiche Verteilungsform (nur mit anderen Kennwerten) wieder übergehen. Ein Beispiel dafür ist die Beta-Verteilung als mögliche Verteilungsform für den Parameter der Verteilung eines dichotomen Merkmals. Beobachtete Daten aus solchen Verteilungen verteilen sich ihrerseits binomial; mit ihren Kennwerten lassen sich wiederum die Parameter der Beta-Verteilung revidieren. Ähnliches gilt für Normalverteilungen und aus ihr abgeleitete Verteilungen.

Wir können hier auf Einzelheiten nicht weiter eingehen; der interessierte Leser findet mehr in Büchern wie z.B. Winkler, 1972; de Groot, 1970; Phillips, 1973; Hays, 1973.

## 6.4 Das Principle of Stable Estimation bei der Parameterschätzung

Das bereits früher angesprochene Prinzip der Robustheit der Schätzung (principle of stable estimation) besagt im Falle der Parameterschätzung, daß man nach Möglichkeit eine Gleichverteilung als a-priori-Verteilung über den Raum mögliche Parameter haben sollte, es genügt aber auch, wenn sie sich in dem von den Daten am meisten begünstigten Bereich nicht sprunghaft verändert und keinen falschen Bereich allzusehr bevorzugt.

## 7. Die Erhebung von a-priori- Wahrscheinlichkeiten

Während sich das Problem der Gewinnung der Wahrscheinlichkeiten der Daten unter Annahme der jeweiligen Hypothesen in der Bayes-Statistik nicht von dem der klassischen Statistik unterscheidet, kommt hier ein weiteres hinzu: die Erhebung der a-priori-Wahrscheinlichkeiten der Hypothesen. Obwohl wir bei der Besprechung des Prinzips der Robustheit der Schätzung (principle of stable estimation) schon an Beispielen gesehen haben, daß Fehler bei der Einschätzung dieser Hypothesenwahrscheinlichkeiten durch die Daten in kürzester Zeit ausgeglichen werden können, wollen wir uns doch noch mit der Frage beschäftigen, wie man denn überhaupt zu möglichst brauchbaren a-priori-Wahrscheinlichkeiten kommt. Die Antwort ist (wie immer, wenn man in der Psychologie etwas nicht genau weiß): Man läßt sie durch Experten schätzen. Um allerdings möglichst genaue Schätzungen zu bekommen, kann man Ver-

fahren anwenden, die auf dem Prinzip der Erwartungsmaximierung aufbauen: die sog. Belohnungsfunktionen oder scoring-rules. Ihre Aufgabe besteht hauptsächlich darin, für den Schätzer einen (motivationalen) Anreiz zu schaffen, sein Fachwissen über die Wahrscheinlichkeit einer Hypothese möglichst genau in eine Zahl zu kleiden. Tut er das, so maximiert er damit den (subjektiven) Erwartungswert seines Einkommens in der betreffenden Schätzsituation.

Dies soll am Beispiel einer einfachen Belohnungsfunktion, der sog. quadratischen scoring rule demonstriert werden. Sei im folgenden  $P$  die (subjektive) Wahrscheinlichkeit des Schätzers für das Eintreten eines zufälligen Ereignisses (in unserem Falle: die Richtigkeit einer Hypothese) und  $X$  die Zahl, die der Schätzer als seine subjektive Wahrscheinlichkeit angibt. Dem Schätzer wird dann eine materielle (Geld) oder ideelle Verteilung von („Punkten“) Belohnung für seine Schätzung gegeben, bei der von einem Festbetrag  $F$  ein Betrag abgezogen wird, der eine quadratische Funktion der Differenz  $(X - P)$  ist; im einfachsten Fall ist die Belohnung für den Schätzer dann:  $F - (X - P)^2$ .

Ohne die Differentialrechnung zu bemühen, ist in diesem Falle leicht zu erkennen, daß der Schätzer seinen Gewinn maximiert, wenn er als seine Schätzung  $X$  genau den Wert  $P$  angibt, weil dann der Wert in der Klammer gleich 0 ist, und somit nichts von  $F$  abgezogen wird.

Die obige Berechnung setzt allerdings voraus, daß zusätzlich zur Schätzung  $X$  auch die „wahre“ subjektive Wahrscheinlichkeit  $P$  bekannt ist - was in der Regel nicht der Fall sein wird, denn sonst wäre die ganze Prozedur ja unnötig, und wir könnten direkt  $P$  verwenden. In leicht modifizierter Form läßt sich die quadratische Belohnungsfunktion aber auch dann anwenden, wenn die Wahrscheinlichkeit  $P$  des betreffenden zufälligen Ereignisses nicht bekannt ist. Wir setzen als Belohnung ein:

$$\begin{aligned} &F - (X-1)^2, \text{ falls das Ereignis eintritt} \\ &F - (X-0)^2, \text{ falls das Ereignis nicht eintritt.} \end{aligned}$$

Da das Ereignis - jedenfalls nach der zu erhebenden Meinung des Schätzers - mit der Wahrscheinlichkeit  $P$  auftreten wird, beträgt sein Erwartungswert

$$\begin{aligned} &P \cdot (F - (X-1)^2) + (1-P) \cdot (F - X^2) \\ &= PF - P(X^2 - 2X + 1) + F - X^2 - PF + PX^2 \\ &= PF - PX^2 + 2PX - P + F - X^2 - PF + PX^2 \\ &= F - P + 2PX - X^2, \end{aligned}$$

den er durch Wahl einer Schätzung  $X$  zu maximieren trachten soll. Wir bilden dazu

$$\frac{d(F - P + 2PX - X^2)}{dX} = 2P - 2X,$$

setzen dieses gleich 0 und erhalten  $X = P$ . Auch bei dieser Belohnungsfunktion maximiert der Schätzer also seinen Erwartungswert, wenn er als Schätzung  $X$  genau die Wahrscheinlichkeit  $P$  angibt.

Die oben beschriebene quadratische Belohnungsfunktion ist nur eine von vielen, die in den vergangenen Jahren entwickelt worden sind. Kompliziertere wie die sog. Rang-Belohnungsfunktionen berücksichtigen bei der gleichzeitigen Schätzung mehrerer Wahrscheinlichkeiten (für mehr als 2 konkurrierende Hypothesen) sogar die Rangfolge von Schätzungen und Wahrscheinlichkeiten (Stäel von Holstein (1977)). Übersichten über Belohnungsfunktionen findet man bei Murphy & Winkler (1970).

Es mag für manchen etwas unglaublich und unnötig erscheinen, schon an sich hinreichend motivierte Wissenschaftler (bei den in Kiel promovierten sogar: vereidigte Wahrheitssucher) mit solchen Belohnungsfunktionen möglichst genaue Wahrscheinlichkeitsschätzungen entlocken zu wollen; sie sollten ohnedies hinreichend darum bemüht sein. Ausgedehnte Versuchsreihen mit Meteorologen bei der Vorhersage von Niederschlägen und Temperaturen (Murphy & Winkler, 1977; Winkler & Murphy, 1968), mit Bankfachleuten und Wirtschaftswissenschaftlern bei der Vorhersage der Entwicklung von Wertpapier- und Devisenkursen (Borcherding, 1978), Studenten bei der Schätzung demographischer Daten (Aschenbrenner & Wendt, 1978) und nicht zuletzt mit dem „Mann von der Straße“ bei der Vorhersage der Ergebnisse von Fußballspielen (de Finetti, 1962) haben aber gezeigt, daß durch die Anwendung von Belohnungsfunktionen erhebliche Verbesserungen von Schätzleistungen möglich sind, daß sie sich auch als Verstärker zur Rückmeldung beim Lernen von Ereigniswahrscheinlichkeiten eignen, ebenso wie zur Bewertung der Güte von Wahrscheinlichkeitsschätzungen. Ihre Anwendung kann daher auch für den Wissenschaftsprozess empfohlen werden, wenn es darum geht, möglichst genaue a-priori-Wahrscheinlichkeiten  $P(H_i)$  für die konkurrierenden Hypothesen  $H_i$  zum Zwecke der Bayes-Revision zu finden. Das Prinzip der Robustheit der Schätzung mag dabei als Trost dienen, daß es auf allzu genaue Schätzungen gar nicht ankommt; schon die ersten auftretenden Daten werden die a-priori-Schätzung rasch in den richtigen Bereich bringen.

Im übrigen sollte man bei der Planung empirischer Untersuchungen ohnedies nach dem Ökonomie-Prinzip vorgehen und versuchen, den Informationsgewinn durch die Untersuchung zu maximieren. Nach den Erkenntnissen der Informationstheorie ist das genau dann der Fall, wenn die konkurrierenden Hypothesen a priori (zumindest subjektiv) gleich wahrscheinlich sind:  $P(H_0) = P(H_1) = 0.5$ . In diesem Falle ist der Likelihoodquotient  $L$  gleich dem Verhältnis der Wahrscheinlichkeiten der Hypothesen a posteriori; die Bayes-Statistik läßt sich in diesem Fall auf den Likelihood-Quotienten-Test (mit anderer Interpretation der Wahrscheinlichkeit) zurückführen.

Kritiker der Bayes-Statistik (und der subjektiven Wahrscheinlichkeits-Interpretation als „Glaubensgrad“ überhaupt) haben gegen den oben beschriebenen Ansatz zur Erhebung von (subjektiven) Wahrscheinlichkeiten mit Hilfe von Belohnungsfunktionen erhebliche Vorbehalte, so z.B. Kemnitz (1980). Ein Teil der Schwierigkeiten kommt daher, daß es zunächst schwer fällt, sich die Wahrscheinlichkeit des Eintretens eines Ereignisses so wie andere Attribute oder Eigenschaften von Gegenständen und/oder Ereignissen als kontinuierliche Variable vorzustellen, über die man genauso wie über andere Eigenschaften aufgrund von Wahrnehmungen und anderen Erfahrungen Urteile abgeben kann. Hinter der Einführung der Belohnungsfunktionen als Lernhilfe beim Wahrscheinlichkeitsschätzen steht die Idee, daß man die Eigenschaft „Wahrscheinlichkeit des Auftretens“ abschätzen kann wie die Höhe oder Klangfarbe eines Tones oder den Rotanteil einer Farbe. Dabei spielt es eine untergeordnete Rolle, ob dieses Lernen auf der Basis von Häufigkeiten des Auftretens genau des fraglichen Ereignisses geschieht (wie z.B. bei Aschenbrenner & Wendt, 1978) oder ob es aufgrund von Analogieschlüssen aus anderen, aber ähnlichen Situationen und Ereignissen stattfindet. Es war vielleicht ungeschickt von manchen Subjektivisten, überhaupt von „Glauben“ und „Gefühl“ zu sprechen, weil diese Begriffe Assoziationen wecken, die hier nicht gemeint sind. Besser wäre es, von Wahrscheinlichkeit als unsicherem Wissen zu reden. Werde ich beispielsweise aufgefordert, eine Aussage über die durchschnittliche Größe eines Maulwurfes zu machen (und ich habe nicht gerade einen zur Hand), so werde ich auf der Basis meiner Erfahrungen mit Maulwürfen einen Mittelwert schätzen und als „Unsicherheit“ dieser Angabe (dieses „Wissens“) dazu noch ein Streuungsmaß. Analog kann ich über die Wahrscheinlichkeit des Auftretens eines Maulwurfes auf einem bestimmten Grundstück eine Schätzung abgeben, und als „Unsicherheit“ dazu die Streuung der entsprechenden Beta-Verteilung. Und beim Lernen von eben solchen Schätzungen aufgrund von Erfahrungen sollen die Belohnungsfunktionen helfen.

## *8. Die Bewertung der Ausgänge von Entscheidungen*

Die bisher besprochenen Verfahren liefern noch keinen Anhaltspunkt dafür, wie man denn sein Kriterium zu wählen habe, nach dem man eine Nullhypothese verwirft bzw. akzeptiert. Wie hoch soll man denn die zulässigen Irrtumswahrscheinlichkeiten  $\alpha$  und  $\beta$  beim klassischen Signifikanztest bzw. Sequenztest wählen, wie groß muß der Likelihoodquotient zugunsten einer Hypothese mindestens sein, oder wie groß die a-posteriori-Wahrscheinlichkeit  $P(H \mid D)$  einer Hypothese nach der Bayes-Revision, ehe man sie für hinreichend bestätigt halten kann? Da die Wahl dieses Kriteriums Einfluß auf die Wahrscheinlichkeiten der möglichen Folgen der Entscheidung hat, sollte sie diese möglichen Folgen in irgendeiner Weise berücksichtigen, etwa derart, daß schwerwiegende und kostspielige Folgen Fehlentscheidungen möglichst

wenig wahrscheinlich gemacht werden, dagegen belanglosere eher in Kauf genommen werden, und wünschenswerte (richtige) Entscheidungen möglichst hohe Wahrscheinlichkeiten erhalten.

Zu dieser Wahl eines Signifikanzniveaus oder einer sonstigen kritischen Wahrscheinlichkeit zur Entscheidung geben die meisten Statistik-Lehrbücher wenig an Hilfe. Lienert (1973) behandelt dieses Problem, indem er zwischen „praktischer“ und „statistischer“ Signifikanz unterscheidet (S. 73). Andere Autoren sprechen hier von dem „semantischen“ Aspekt der Signifikanzprüfung, dem sie den rein formalen Algorithmus des Tests als „syntaktischen“ Aspekt gegenüberstellen.

Mit diesem Problem inhaltlich („semantisch“) eng verbunden, aber formal abzutrennen ist die Frage, wie groß denn eine „Wirkung“ einer unabhängigen Variablen auf eine abhängige sein muß (oder - äquivalent - wie groß ein Unterschied zwischen zwei Gruppen, oder ein korrelativer Zusammenhang zwischen zwei Variablen sein muß), um als berichtenswert anerkannt zu werden, d.h. als bedeutsam genug, um als Satz in den Kanon des „Wissens“ aufgenommen zu werden, der die Wissenschaft ausmacht.

Es geht hier also um den Abstand der konkurrierenden Hypothesen, oder genauer: der Parameter, über welche die Hypothesen etwas aussagen. Wie wir bei der Besprechung der Frage der Stichprobengröße gesehen haben, geht das Quadrat dieses Abstandes in den Nenner der Formel zur Berechnung des erforderlichen Stichprobenumfanges ein - das bedeutet, daß praktisch jeder beliebig kleine Unterschied oder Zusammenhang durch Wahl einer hinreichend großen Stichprobe bei jeder beliebig kleinen Irrtumswahrscheinlichkeit „signifikant“ gemacht werden kann - zumindest statistisch.

In diesem Zusammenhang sei ausdrücklich darauf hingewiesen, daß die erreichte Höhe des Signifikanzniveaus an sich noch gar nichts über die Stärke des gefundenen Effektes aussagt. Dieses Mißverständnis findet man aber bei den Benutzern klassischer statistischer Testverfahren häufig. Als ein Beispiel sei Melton (1962) erwähnt, der glaubte, daß „signifikantere“ Ergebnisse - also solche bei einem niedrigeren  $\alpha$  - sich leichter reproduzieren lassen müßten als solche bei einem höheren  $\alpha$ . Tatsächlich ist aber das Gegenteil der Fall, denn unter sonst gleichen Bedingungen und bei Richtigkeit von  $H_1$  wird bei kleinerem  $\alpha$  auch  $(1 - \beta)$  kleiner, damit sinkt also die Wahrscheinlichkeit der Reproduktion des Ergebnisses mit  $\alpha$ , statt sich zu erhöhen (Witte, 1980; vgl. auch Hagen & Seifert, 1979).

Zu dem Irrglauben, das erreichte Signifikanzniveau könne etwas mit der Stärke des beobachteten Effektes oder Zusammenhanges zu tun haben, trägt wohl auch viel der (schon rein sprachlich) unschöne Brauch bei, Ergebnisse auf dem 5%-Verlässlichkeitsniveau als „signifikant“, solche auf dem 1 %-Niveau als

„sehr signifikant“ oder „hochsignifikant“, und solche auf dem 0,1%-Niveau gar als „höchstsignifikant“ zu bezeichnen.

Um über die Stärke von Effekten oder Zusammenhängen eine sinnvolle Aussage zu machen, wäre es im Rahmen der klassischen Statistik besser, anstelle des erreichten Signifikanzniveaus den Anteil der Varianz einer Variablen zu berichten, der von der anderen Variablen kontrolliert (festgelegt) wird, bei linearen korrelativen Zusammenhängen also den Determinationskoeffizienten  $r^2$ , bei nichtlinearen  $\eta^2$ , bei varianzanalytischen Versuchsplänen das Effektivitätsmaß  $\omega^2$ , wie bereits weiter oben angesprochen. (Näheres hierzu s. z.B. auch Hofstätter & Wendt, 1974, Kap. 17.)

Es ist daher wichtig, bereits vor Durchführung der Untersuchung eine Entscheidung darüber zu treffen, was man denn eigentlich entdecken will oder für entdeckenswert erachtet, genauer: ab welcher numerischen Größe eines Unterschiedes oder Zusammenhanges man diesen für praktisch oder theoretisch bedeutsam und mitteilens- und wissenswert hält. Selbst wenn man keine präzise Alternativhypothese (z. B.  $H_1 : \Theta = \Theta_1$ ) aufstellt, sollte man doch zumindest den Bereich angeben, in dem man die Alternativhypothese als hinreichend von der Nullhypothese verschieden ansieht, um sie wirklich als „Alternative“ betrachten zu können. Anstelle des Hypothesenpaares

$$\begin{aligned} H_0: \Theta &= \Theta_0 \\ H_1: \Theta &\neq \Theta_0 \end{aligned}$$

sollte man zumindest eine Alternativhypothese aufstellen, die sich bei ungerichteter Fragestellung etwa schreiben ließe als

$$H_1: \Theta \quad \begin{cases} \leq \Theta_0 - \Delta \\ \text{oder} \\ \geq \Theta_0 + \Delta \end{cases}$$

Darin ist  $\Delta$  der kleinste noch erwähnenswert erscheinende Unterschied (gegenüber  $\Theta_0$ ), deren Größe allerdings von inhaltlichen Erwägungen abhängt. In industriellen, ökonomischen und technischen Anwendungen ist es oft möglich, dieses  $\Delta$  über Kosten-Nutzen-Analysen oder ähnliche Überlegungen festzulegen; beispielsweise kann man herausfinden, wie groß die Wirkung  $\Delta$  einer Änderung eines Fertigungsprozesses sein muß, wenn sie sich in einer gegebenen Zeit amortisieren soll, und man kann dies dann den Investitionskosten gegenüberstellen. Für Pharmaka und Therapien in der Medizin, für Curricula in der Pädagogik, gelegentlich auch für soziale Hilfsprogramme lassen sich ähnliche Überlegungen anstellen, aber in der sog. „reinen“ Forschung dürfte das schwieriger sein. Trotzdem sollte auch der „reine“ Forscher schon bei der Planung seiner Vorhaben überlegen, ab welcher Größenordnung er seine möglichen Befunde als relevant für seine Fragestellung betrachten würde - beispielsweise wenn ein anderer sie entdeckt hätte.

Ist nun (nach den obigen Überlegungen) eine Wahl von Null- und Alternativhypothese getroffen, so muß zur weiteren Anwendung entscheidungstheoretischer Strategien zunächst eine Bewertung der Folgen oder Ausgänge der Entscheidung vorgenommen werden. Nehmen wir zunächst wieder den einfachen Fall zweier konkurrierender Hypothesen (und zweier möglicher „wahrer“ Zustände der Natur, die ihnen entsprechen), so haben wir vier mögliche Ausgänge der Entscheidungssituation, die in Tabelle 2 dargestellt sind.

Tabelle 2:

Mögliche Ausgänge einer Entscheidungssituation		Wahr ist	
		$H_0$	$H_1$
Entschieden wird für	$H_0$	III	II
	$H_1$	I	IV

Darin sind die Ausgänge (Zellen der Matrix) I und II wieder die bereits aus der klassischen Testtheorie bekannten Fehler I. und II. Art (I: Entscheidung für  $H_1$  bei Zutreffen von  $H_0$ , und II: Entscheidung für  $H_0$  bei Zutreffen von  $H_1$ ), während III und IV die „richtigen“ Entscheidungen zugunsten von  $H_0$  und  $H_1$  repräsentieren.

Es ist plausibel, daß uns die richtigen Entscheidungen wünschenswerter als die Fehlentscheidungen erscheinen sollten, also die Werte der Ausgänge III und IV höher als die der Ausgänge I und II.

Darüber hinaus ist es denkbar, daß einem Wissenschaftler die richtige Bestätigung seiner Forschungshypothese  $H_1$ , also der Ausgang IV, wertvoller erscheint als die (ebenfalls richtige) Beibehaltung der Nullhypothese  $H_0$ , daß leider kein Zusammenhang entdeckbar, sondern alles Beobachtete im Bereich des Zufalls liegt.

Unter den beiden möglichen Fehlentscheidungen I und II ist ihm möglicherweise II lieber als I: lieber einen Zusammenhang noch nicht entdecken, als irgendwelche Zufälligkeiten als neue Erkenntnisse in die Welt hinausposaunen, jedenfalls neigen die meisten Wissenschaftler eher zu dieser mehr „konservativen“ Haltung. Praktisch äußert sich das in ihrer Wahl von  $\alpha$  und  $\beta$  mit  $\alpha < \beta$ .

Wir haben hier durch ein paar einfache Überlegungen, ohne irgendeinen Allgemeinheitsanspruch damit zu verknüpfen, schon eine Rangordnung der Be-



Wertungen für die vier möglichen Ausgänge der Entscheidungssituation gefunden:  $I < II < III < IV$ .

(Diese Rangfolge muß keineswegs in jeder Situation die gleiche sein.)

Damit ist der erste Schritt zu einer Skalierung der Werte der einzelnen Ausgänge der Entscheidung getan. Wir können diese aber noch weiter verfeinern. Dazu bitten wir den Wissenschaftler (oder die Institution), der die Entscheidung treffen soll (oder von ihr und ihren Folgen betroffen ist), sich folgendes vorzustellen:

Es sei eine Wahl zwischen drei von den vier möglichen Ausgängen zu treffen, und zwar zwischen dem besten oder schlechtesten Ausgang, oder einem der dazwischenliegenden. Natürlich wird in dieser Situation jeder den besten Ausgang haben wollen - in unserem Beispiel Nummer IV -, aber den bekommt er nicht so ohne weiteres. Zwischen dem besten und dem schlechtesten Ausgang wird nämlich eine Lotterie veranstaltet, wogegen man den dazwischenliegenden Ausgang mit Sicherheit bekommen kann, wenn man ihn wählt. Die Wahrscheinlichkeit  $Q$ , mit der man bei der Lotterie zwischen dem besten und dem schlechtesten Ausgang den besten bekommt, wird nun so lange variiert, bis der Entscheidungsträger zwischen der sicheren mittleren Alternative und der Lotterie zwischen der besten und schlechtesten indifferent ist.

Die Wahrscheinlichkeit  $Q$ , bei der das der Fall ist, dient dann als Skalenwert der jeweiligen sicheren Alternative; die beste und schlechteste werden mit 1 und 0 bewertet.

In unserem Beispiel müßten auf diese Weise die beiden möglichen mittleren Ausgänge II und III skaliert werden: Der Entscheidungsträger muß jeweils entscheiden zwischen der mittleren Alternative (II bzw. III) und einer Lotterie, bei der er mit Wahrscheinlichkeit  $Q$  ( $Q_{II}$  bzw.  $Q_{III}$ ) den bestmöglichen Ausgang IV oder mit Wahrscheinlichkeit  $1-Q$  den schlechtesten Ausgang I erhält. Dabei wird  $Q$  jeweils etwas erhöht, bis ihm die Lotterie und die sichere mittlere Alternative gleich wünschenswert erscheinen. Praktisch ist das der gleiche Vorgang wie bei der Schwellenbestimmung in der Psychophysik. Theoretisch baut diese Methode der Nutzenmessung auf der Axiomatik von von Neumann & Morgenstern (1944, 2. Aufl. 1947) auf. Näheres hierzu s. Keeney & Raiffa (1976).

Wir haben das Verfahren hier für den einfachsten Fall mit 2 Hypothesen mal 2 Zuständen = vier möglichen Ausgängen dargestellt; es läßt sich aber ebenso gut bei mehr Ausgängen anwenden. In jedem Fall werden alle dazwischenliegenden Alternativen einzeln mit einer Lotterie zwischen der besten und der schlechtesten verglichen und die Wahrscheinlichkeit in dieser Lotterie so lange verändert, bis Indifferenz besteht. Wir haben damit ein Verfahren zur Bestimmung von Nutzwerten für die möglichen Ausgänge einer Entscheidung vorge-

stellt, das sich auch anwenden läßt, wenn keinerlei „objektive“ Gesichtspunkte für die Bewertung der Alternativen gegeben sind, sondern nur von den Präferenzen des Entscheidungsträgers ausgegangen werden kann.

Falls irgendwelche ökonomischen Gesichtspunkte bestehen, die zur Bewertung der möglichen Ausgänge der Entscheidungssituation beitragen können, so sollten diese mit herangezogen werden. Bei Entscheidungen im Bereich von Industrie und Handel ist es oft möglich, direkt monetäre Kosten und Nutzen für die verschiedenen Ausgänge anzugeben.

## 8.1 Bewertung multiattributiver Ausgänge

In komplexen Entscheidungssituationen sind die Wahlalternativen oder - in unserem Falle - Ausgänge der Entscheidung oft durch eine Vielfalt von Eigenschaften charakterisiert, die alle mehr oder weniger zur Bewertung der Alternative beitragen.

Obwohl dieses Kapitel primär für die Situation des Forschers geschrieben ist, der angesichts beobachteter Daten über die Annahme oder Verwerfung bestimmter Wissens-Sätze entscheiden soll, so wollen wir hier zur Illustration doch ein Beispiel herausgreifen, das weniger forschungsspezifisch ist, aber zu dem sich jeder aufgrund seiner Alltagserfahrung konkret etwas vorstellen kann. Sucht jemand beispielsweise eine Wohnung oder ein Zimmer zu mieten oder ein Haus zu kaufen, so ist eine Fülle von Gesichtspunkten zu berücksichtigen, wie Lage, Größe, Preis, und viele Details der Ausstattung. Zunächst einmal ist es wichtig, zu erheben, welche solcher Gesichtspunkte überhaupt für die Entscheidung relevant sind. Diese Gesichtspunkte, Aspekte, Merkmale, Attribute oder auch „Kriterien“ (wie sie manchmal genannt werden) der Alternativen liegen teilweise bereits als zahlenmäßige Größen fest (wie Preis, Größe in Quadratmeter usw.) (aber nicht deren Nutzen), teilweise müssen sie erst skaliert werden, ehe sie weiterer quantitativer Verarbeitung zugänglich sind. Dazu bieten sich die herkömmlichen Skalierungsmethoden an, ebenso auch das oben beschriebene Lotterie-Äquivalent-Verfahren. So erreichen wir schließlich eine zahlenmäßige Repräsentation aller entscheidungsrelevanten Attribute jeder Alternative - sei es aufgrund gegebener Maße, sei es durch die Anwendung von Skalierungsverfahren -, so daß sich jede Alternative als Vektor der Ausprägungen ihrer Attribute darstellen läßt. Aus diesen müssen wir nun wieder zu einer globalen Bewertung der Alternativen kommen. Wir brauchen dazu eine Funktion, welche den Vektor der Ausprägungsgrade der Attribute jeder Alternative in eine reelle Zahl, eben den Gesamtwert der Alternative, abbildet. Die einfachsten Funktionen, die das tun und die für solche Fälle auch am häufigsten herangezogen werden, sind die additive und die multiplikative, bei der die einzelnen Maße der Attribute (Komponenten des

o.a. Vektors) mit Gewichtungsfaktoren multipliziert und diese Produkte dann über die Attribute (Komponenten) addiert oder multipliziert werden:

$$U_i = \sum_j w_j u_j(x_{ij}) \quad \text{oder} \\ V_i = \prod_j [1 + k w_j u_j(x_{ij})]^{1/k} - 1/k$$

(n. Keeney & Raiffa, 1976)

worin  $U_i$  bzw.  $V_i$  den Gesamtwert der Alternative oder des Ausgangs  $i$  bezeichnet,  $x_{ij}$  das Ausmaß, in dem die Alternative  $i$  (das Attribut)  $j$  besitzt  $u_j(x_{ij})$  deren Nutzen, und  $w_j$  das Gewicht, das der Entscheidungsträger diesem Attribut  $j$  zumißt. Die relative Größe der Gewichte  $w_j$  macht dabei gleichzeitig auch die verschiedenen zur Messung der Attribute verwendeten Skalen vergleichbar.

Die multiplikative Form der Verknüpfung ( $V_i$ ) hat gegenüber der additiven Form ( $U_i$ ) den Vorteil, daß hier ein Nullwert auf einem Attribut den Gesamtwert der Alternative Null werden lassen kann, während bei der additiven Verknüpfung auch bei einem Nullwert auf einem Attribut immer noch ein beträchtlicher Gesamtwert aus den anderen Attributen zusammenkommen kann. Es ist bei der Wahl einer Verknüpfungsfunktion zu überlegen, ob das sinnvoll ist.

Beide Verknüpfungsfunktionen nehmen an, daß die Attribute unabhängig voneinander zum Gesamtwert beitragen; diese Unabhängigkeitsforderung ist vergleichbar der des additiven Modells der Varianzanalyse oder der additiv verbundenen Messung (additive conjoint measurement).

Weiterhin setzen sie beide voraus, daß jedes Attribut isoton zum Gesamtwert beiträgt (monoton, ggf. auch antiton mit negativem  $w_j$ ) und nicht etwa irgendwo im endlichen Bereich ein Optimum hat, von dem aus der Erwünschtheitsgrad des Attributs nach beiden Seiten sinkt. Bei dem Beispiel von der gesuchten Wohnung könnte das beispielsweise der Fall sein: Man hat eine Vorstellung von einer idealen Größe (qm-Zahl) der Wohnung, vielleicht einen idealen Bereich, aber sowohl zu große wie auch zu kleine Wohnungen werden weniger erstrebenswert. Falls für einzelne Attribute solche eingipfeligen (oder auch mehrgipfeligen) Präferenzfunktionen bestehen, muß diesem Umstand bei der Skalierung des Attributs Rechnung getragen werden; beispielsweise kann man als Skalenwert den Abstand der jeweiligen Alternative vom Idealpunkt auf diesem Attribut verwenden.

Um dies an unserem Wohnungs-Beispiel konkret werden zu lassen: Schwebt jemandem beispielsweise eine ideale Wohnungsgröße von 80 qm vor, so wird man nicht die Größe in qm als Maß des Attributs „Wohnungsgröße“ in die Nutzenberechnung eingehen lassen, sondern etwa  $x_{ij} = |qm_i - 80|$  (mit entsprechend negativem  $w_j$ ).

Geht es in einer Entscheidungssituation nicht um die Aufstellung einer allgemeingültigen Bewertungsfunktion (die auch für künftige Fälle verfügbar sein soll), sondern nur um die Bewertung konkret vorliegender Alternativen oder Ausgänge, so kann man auch attributweise wieder nach der im vorigen Abschnitt beschriebenen Methode vorgehen: Rangordnung aller Alternativen (aber nun nur unter Berücksichtigung des zu skalierenden Attributs), Bewertung der auf diesem Attribut besten und schlechtesten Alternative mit 100 (bzw. 1) und 0, Konstruktion von Lotteriespielen, bei denen der Befragte jeweils drei konstruierte Alternativen vorgelegt bekommt, die sich in allen Attributen gleichen bis auf die zu skalierende, wobei er zwischen der zu skalierenden Alternative einerseits und andererseits einer Lotterie wählen muß, bei der er mit der Wahrscheinlichkeit  $x_{ij}$  die beste und mit der Wahrscheinlichkeit  $(1 - x_{ij})$  die schlechteste bekommt.  $x_{ij}$  wird solange variiert, bis der Befragte indifferent zwischen der sicheren Alternative und der Lotterie ist.

Zur Ermittlung der Gewichte  $w_j$  kann dann das gleiche Verfahren angewendet werden: Die Attribute (und damit ihre Gewichtungsfaktoren  $w_j$ ) werden zunächst in eine Rangordnung nach ihrer Wichtigkeit gebracht. Dann werden wieder drei konstruierte Alternativen A, B und C vorgelegt, die sich in allen Attributen gleichen bis auf drei: Alternative A hat auf dem wichtigsten Attribut die Ausprägung 100, auf dem unwichtigsten und dem zu skalierenden Attribut dagegen die Ausprägung 0. Alternative B hat auf dem unwichtigsten Attribut die Ausprägung 100 und auf dem wichtigsten und dem zu skalierenden die Ausprägung 0, und Alternative C hat auf dem wichtigsten und auf dem unwichtigsten Attribut jeweils die Ausprägung 0, aber auf dem zu skalierenden Attribut die Ausprägung 100. Der Befragte hat dann die Wahl zwischen der sicheren Alternative C und einer Lotterie, bei der er mit der Wahrscheinlichkeit  $w_j$  die Alternative A, und mit der Wahrscheinlichkeit  $(1 - w_j)$  die Alternative B bekommt,  $w_j$  wird wieder so lange variiert, bis der Befragte indifferent zwischen der sicheren Alternative C und der Lotterie ist.

Die additive und die multiplikative Verknüpfungsfunktion zur multiattributiven Nutzenmessung setzen weiterhin voraus, daß die verschiedenen Attribute kompensatorisch füreinander eintreten können: Ist eine Alternative auf einem ihrer Attribute weniger attraktiv für den Entscheidungsträger als die andere, so kann dies durch größere Attraktivität auf einem anderen Attribut ausgeglichen werden. Es gibt aber auch Entscheidungssituationen, in denen zunächst nur das wichtigste Attribut beachtet wird und das zweitwichtigste Attribut erst zum Tragen kommt, wenn zwei Alternativen auf dem wichtigsten Attribut gleichwertig erscheinen. Dies läßt sich fortsetzen: das drittwichtigste Attribut wird erst dann berücksichtigt, wenn Alternativen auf dem ersten und zweiten gleichwertig erscheinen. Hier kann keine Kompensation der Attribute füreinander eintreten. Man spricht in diesem Falle von einer hierarchischen Struktur der Attribute. Hierarchische und kompensatorische Verknüpfungsfunktionen

lassen sich aber kombinieren, z.B. kann sich eine Menge von Alternativen zunächst nach dem wichtigsten Attribut ohne Berücksichtigung der anderen ordnen lassen, bei Gleichheit von Alternativen auf diesem ersten Attribut kann dann aber eine kompensatorische Kombination weiterer Attribute für die weitere Beurteilung der Attraktivität herangezogen werden.

Die oben beschriebenen Verfahren zur Ermittlung der Bewertung oder des Nutzens von Wahlalternativen oder Ausgängen von Entscheidungen mögen recht kompliziert klingen; in der Tat kostet es einige Zeit und Mühe, einen Entscheidungsträger oder andere daran interessierte Personen mit dem Verfahren vertraut zu machen und dann auf diese Weise seine Nutzwerte zu ermitteln. Es tröstet wahrscheinlich auch wenig, zu erfahren, daß man hierfür Computer-Dialoge programmieren kann. Für die Anwendung dieser Strategien in Wirtschaft und Politik hat sich im Laufe des letzten Jahrzehnts sogar ein neuer Berufszweig entwickelt, der des „decision analyst“ (zu Deutsch vielleicht „Entscheidungsanalytiker“). Seine Tätigkeit besteht darin, als Berater den Verantwortlichen im Top-Management von Firmen, Regierungen und Verwaltungsorganisationen ihre Entscheidungen in komplexen Situationen durch Anwendung solcher Verfahren zu erleichtern. Seine Bemühungen, aus seinen Klienten mit einer Vielfalt von Erhebungstechniken wie den oben beschriebenen zunächst herauszubekommen, was er „eigentlich“ will, und wie sich seine Wahlalternativen durch Attribute und Ausprägungen auf diesen charakterisieren lassen, ähneln oft mehr denen eines Psychotherapeuten als denen eines Wissenschaftlers.

Trotz dieser überwiegend kommerziell ausgerichteten Aktivität der Decision Analysis ist die „reine“ Wissenschaft von ihr nicht unberührt geblieben: Nicht nur bei der NASA helfen decision analysts bei der Festlegung der Forschungsprogramme, sondern auch in den meisten größeren amerikanischen Institutionen, die Forschungsprojekte in anderen Gebieten fördern (s. z.B. Edwards, Guttentag & Snapper, 1975). Was dort zur Verteilung und zum sinnvollen Einsatz der verfügbaren Mittel notwendig geworden ist, kann sich auch der einzelne Forscher oder das einzelne Institut bei Entscheidungen über den Einsatz seiner eigenen Mittel zunutze machen. Man sollte auf jeden Fall zunächst versuchen, mit dem im vorigen Abschnitt beschriebenen Verfahren eine globale Bewertung der Wahlalternativen oder der möglichen Ausgänge einer Entscheidung zu erreichen. Es sind aber Fälle denkbar, in denen dies nicht gelingt - was sich u.a. durch Intransitivitäten in der Bevorzugungsranordnung und andere Inkonsistenzen bemerkbar macht. Dann liegt dieses Versagen des einfachen Verfahrens häufig daran, daß die Vielfalt der zu berücksichtigenden Aspekte (Attribute) die Informationsverarbeitungskapazität des Befragten (Entscheidungsträgers oder Betroffenen) überfordert. Eine Aufgliederung der Bewertungsaufgabe in Einzelaspekte in der oben beschriebenen Art führt dann zum Ziel.

Die Praxis hat dabei gezeigt, daß sehr wichtig für das Gelingen einer solchen „multiattributiven Nutzenanalyse“ ist, die zu skalierenden Attribute adäquat auszuwählen, zu formulieren und operationalisieren, am besten unter Mitwirkung des zu Befragenden (z.B. Aschenbrenner, 1977; Dyer & Miles, 1976).

Die obigen Anwendungsbeispiele waren überwiegend Entscheidungen über Forschungsprojekte oder die Bearbeitungswürdigkeit von Fragen, nicht über die Gültigkeit von Hypothesen. Diese ist aber prinzipiell nicht anders zu behandeln: Die Bewertung (des Nutzens) der Förderungswürdigkeit eines Projektes impliziert ja auch die Bewertung des Wissens, das mit ihm gewonnen werden soll, also der Hypothesen, die in ihm bestätigt oder verworfen werden sollen. Die meisten dieser Entscheidungen über die Förderung von Forschungsprojekten entsprechen formal am ehesten der Entscheidung beim Sequenztest: Man muß wählen, ob man die anstehende Fragestellung in der einen oder anderen Richtung aufgrund des bisherigen Wissens als „entschieden“ betrachten, oder ob man „weiterbeobachten“, also das beantragte Forschungsprojekt bewilligen will.

## 9. Entscheidungskriterien

Im folgenden wollen wir annehmen, daß wir auf die oben beschriebene oder irgendeine andere sinnvolle Weise zu Bewertungen für die möglichen Ausgänge oder Folgen einer Entscheidung gekommen sind. Um mehrere Entscheidungskriterien demonstrieren zu können, nehmen wir als Beispiel eine Matrix, in der die 9 Bewertungen der Folgen der Entscheidung zwischen drei konkurrierenden Hypothesen über drei mögliche Zustände der Natur eingetragen sind.

Oben links in dieser Matrix erkennen wir die bereits bekannte von Tabelle 2 wieder, nur sind die möglichen Ausgänge (Zellen) jetzt mit Werten versehen. Rechts und unten sind eine Spalte bzw. Zeile für eine dritte konkurrierende Hypothese  $H_2 : \Theta = \Theta_2$  hinzugekommen. Die Werte sind alle so skaliert, daß der höchste Wert 100 und der niedrigste 0 beträgt.

Um das Ganze etwas zu konkretisieren, kann man z.B. versuchen, sich vorzustellen, daß zur Verbesserung eines bestimmten Zustands -beispielsweise der Jugendkriminalität in einem Bezirk, oder der Unfallträchtigkeit einer Kreuzung - zwei einander ausschließende Aktionsprogramme oder Therapien vorgeschlagen worden sind:  $H_0$  besagt, daß keines der beiden hilft, sondern beide eher zusätzliche Probleme schaffen,  $H_1$  besagt, daß nur das Programm 1 hilft, und  $H_2$  behauptet Entsprechendes von Programm 2. In den Zellen der stark umrandeten Matrix der Tabelle 3 steht, wie man die Folgen einer Entscheidung bewertet, bei der man das Programm der betreffenden Zeile gewählt hat, während das der entsprechenden Spalte das richtige gewesen wäre.

Tabelle 3:

	Bewertung der Ausgänge bei den möglichen Zuständen in der Wirklichkeit (payoffs)			Zeilen- mini- mum der payoffs	Bewertungen ab- züglich des Spaltenmaximum (regrets)			Zeilen- mini- mum der regrets	Summe der Be- wer- tungen	Erwar- tungs- wert
	$\Theta_0$	$\Theta_1$	$\Theta_2$		$\Theta_0$	$\Theta_1$	$\Theta_2$			
und Ent- scheidung $H_0: \Theta = \Theta_0$	III 60	II 20	20	20	0	-80	-60	-80	100	24
	I 0	IV 100	10	0	-60	0	-70	-70	110	63
	30	10	80	10	-30	-90	0	-90	120	33
Spalten-Maximum	60	100	80							
Wahrscheinlichkeit des Zustands, $P(\Theta = \Theta_i   D)$	0.1	0.6	0.3							

Wir wie wir sehen werden, können wir auf die Bewertungen der Ausgänge der Entscheidung (in dem stark umrandeten Teil der Tabelle 3) bereits eine Reihe von durchaus sinnvollen und ernstzunehmenden Entscheidungskriterien anwenden, ohne daß es dazu überhaupt der Berechnung irgendwelcher Wahrscheinlichkeiten bedarf. Wir werden im folgenden an unserer Beispielmatrix eine Reihe rationaler Entscheidungskriterien demonstrieren, die z. B. bei Schneeweiß (1966) oder Wendt (1970) ausführlicher diskutiert werden.

**Walds Minimax-Kriterium:** Will man bei seiner Entscheidung äußerste Vorsicht walten lassen, so kann man beschließen, auf jeden Fall die Alternative zu wählen, bei der der Schaden (im Falle einer Fehlentscheidung oder überhaupt im ungünstigsten Fall) möglichst gering ist. Um dies zu erreichen, sucht man zunächst in jeder Zeile (d.h., bei jeder Wahlalternative) den ungünstigsten Wert auf. Diese sind rechts von der Matrix der Bewertungen in der Spalte „Zeilenminimum“ noch einmal zusammengestellt. Man wählt dann die Alternative (Zeile), bei der dieses Minimum das größte ist (verglichen mit den anderen Zeilen). Etwas formaler: man sucht zunächst in jeder Zeile das Minimum, und wählt dann das maximale Minimum - daher der Name dieses Kriteriums: Minimax. In unserem Beispiel der Tabelle 3 führt uns dieses Kriterium zur Entscheidung für  $H_0$  (1. Zeile).

**Savages und Niebans' Minimax-regret-Kriterium:** Die verschiedenen Zustände der Natur (Spalten der Bewertungsmatrix) erlauben ihrerseits auch bei Wahl der optimalen Alternative (Zeile) unterschiedliche Höchstwerte. Andererseits hat der Entscheidungsträger auf die Zustände (Spalten) keinen Einfluß. Bei einer gegebenen Spalte kann er nur innerhalb dieser durch Wahl der günstigsten Alternative (Zeile) die Folgen optimieren. Diesem Umstand versucht das Minimax-regret-Kriterium Rechnung zu tragen, indem es nur den Teil der Bewertungen berücksichtigt, auf den der Entscheidungsträger durch seine Wahl Einfluß nehmen kann. Um das zu erreichen, wird zunächst von jeder Bewertung das Spaltenmaximum abgezogen. Was bleibt, ist der Verlust, den der Entscheidungsträger durch Wahl einer nichtoptimalen Alternative erleidet. (Bei Wahl der optimalen Alternative ist dieser Verlust 0.) Auf diese Weise entsteht aus der Matrix der Bewertungen (payoffs) die Matrix der „regrets“, in Tabelle 3 rechts neben der Spalte der Zeilen-Minima der ursprünglichen Bewertungen.

Auf diese Matrix wird nun das Minimax-Kriterium angewendet. Wieder werden die Zeilen-Minima aufgesucht (Spalte rechts der regret-Matrix in Tabelle 3), und unter diesen dann die Zeile (Alternative) mit dem größten Zeilenminimum ausgewählt. Bei unserer Beispiel-Matrix führt dieses Kriterium zur Entscheidung für die Hypothese  $H_2$  (3. Zeile).

Laplace's **Summen-Kriterium:** Laplace empfahl, die Summe der Bewertungen zu betrachten, und die Alternative mit der größten Summe zu wählen. Wir



können dieses Kriterium eigentlich nicht guten Gewissens weiterempfehlen und erwähnen es hier nur aus historischen Gründen sowie zur Überleitung zu einigen weiteren Kriterien, bei denen ebenfalls Summen innerhalb der Zeilen gebildet werden, also zur Beurteilung der Alternativen die Bewertungen ihrer Konsequenzen bei mehreren Zuständen berücksichtigt werden.

*Hurvicz' Optimismus/Pessimismus-Kriterium:* Das oben besprochene Waldsche Minimax-Kriterium geht in gewisser Weise von einem sehr pessimistischen Ansatz aus: Es rechnet bei der Betrachtung jeder Alternative mit dem ungünstigsten Ausgang, und sucht dann die Alternative, bei der dieser ungünstigste Ausgang noch der vorteilhafteste ist. Ebenso gut wäre es möglich, von einem optimistischen Ansatz aus die Alternativen zu bewerten und von ihnen nach dem günstigsten möglichen Ausgang zu beurteilen. (Ein solches Kriterium würde bei unserer Beispiel-Matrix zur Wahl von  $H_1$  führen.) Hurvicz hat nun ein Kriterium vorgeschlagen, in dem dieser optimistische Ansatz mit dem Waldschen pessimistischen verknüpft wird, indem bei jeder Alternative der günstigste Ausgang mit einem „Optimismus-Parameter“  $\gamma$  und der ungünstigste Ausgang mit einem entsprechenden „Pessimismus-Parameter“  $(1 - \gamma)$  gewichtet und über diese beiden die Summe gebildet wird. Nehmen wir etwa  $\gamma = 0.6$ , so erhalten wir in unserer Beispiel-Matrix der Tabelle 3:

$$\text{Für } H_0 : 0.6 \cdot 60 + 0.4 \cdot 20 = 44,$$

$$\text{für } H_1 : 0.6 \cdot 100 + 0.4 \cdot 0 = 60,$$

$$\text{für } H_2 : 0.6 \cdot 80 + 0.4 \cdot 10 = 52.$$

Bei einem Optimismus-Parameter  $\gamma = 0.6$  würden wir uns hier also für  $H_1$  entscheiden.

Es ist wichtig, zu beachten, daß dieser Parameter  $\gamma$  nicht mit der Wahrscheinlichkeit eines Zustandes verwechselt werden darf, da er je nach Bewertung der Ausgänge mal den einen, mal einen anderen Zustand gewichtet.

*Bayes-Erwartungsmaximierungs-Kriterium:* Gewichten wir die Ausgänge bei jeder Alternative mit der Wahrscheinlichkeit des jeweiligen Zustandes (Spalte) und summieren wir diese über die Spalten innerhalb jeder Zeile auf, so kommen wir zum Erwartungswert der Alternative. Nach dem Erwartungsmaximierungskriterium soll man sich für diejenige Alternative entscheiden, die den größten Erwartungswert hat. Bei unserer Beispiel-Matrix haben wir angenommen, daß die Zustände  $\Theta_0$ ,  $\Theta_1$  und  $\Theta_2$  mit den Wahrscheinlichkeiten 0.1, 0.6 und 0.3 auftreten können.

Bei unserer Beispiel-Matrix ergibt sich

$$\text{für } H_0 : 0.1 \cdot 60 + 0.6 \cdot 20 + 0.3 \cdot 20 = 24,$$

$$\text{für } H_1 : 0.1 \cdot 0 + 0.6 \cdot 100 + 0.3 \cdot 10 = 63,$$

$$\text{für } H_2 : 0.1 \cdot 30 + 0.6 \cdot 10 + 0.3 \cdot 80 = 33.$$

Bei Anwendung des Erwartungsmaximierungs-Kriteriums würden wir uns also für  $H_1$  entscheiden.

Die Wahrscheinlichkeiten der Zustände sind dabei aber nichts anderes als die Wahrscheinlichkeiten der Hypothesen  $P(H_0) = P(\Theta = \Theta_0)$ ,  $P(H_1) = P(\Theta = \Theta_1)$  und  $P(H_2) = P(\Theta = \Theta_2)$ , denn die Hypothesen sind ja nichts anderes als unsichere, d.h. nur mit gewissen Wahrscheinlichkeiten zu treffende Aussagen über die möglichen Zustände der Natur. Haben wir diese Hypothesen-Wahrscheinlichkeiten aufgrund beobachteter Daten revidieren und verbessern können, so werden wir selbstverständlich diese revidierten Hypothesen-Wahrscheinlichkeiten  $P(H_i | D)$  für die Berechnung der Erwartungswerte einsetzen.

Hier schließt jetzt die Anwendung entscheidungstheoretischer Kriterien mit differenzierter Bewertung der Ausgänge der Entscheidungen an die vorher besprochenen statistischen Verfahren zur Berechnung von Hypothesenwahrscheinlichkeiten  $P(H_i | D)$  an.

Zur Ausnutzung möglichst vieler Informationsquellen bei der Entscheidung zwischen konkurrierenden Hypothesen nach dem Bayes-Verfahren empfiehlt sich damit letztlich folgendes Vorgehen:

- 1) Abschätzung der a-priori-Wahrscheinlichkeiten  $P(H_i)$  der Hypothesen,
- 2) Bewertung der möglichen Ausgänge (Folge) der Entscheidung (Aufstellung der payoff-Matrix),
- 3) Erhebung der Daten  $D$  und Revision der a-priori-Wahrscheinlichkeiten  $P(H_i)$  in a-posteriori-Wahrscheinlichkeiten  $P(H_i | D)$ ,
- 4) Berechnung der Erwartungswerte für die Hypothesen,
- 5) Entscheidung für die Hypothese mit dem größten Erwartungswert nach dem Erwartungsmaximierungskriterium.

## 10. *Schlußbemerkung*

Ich bin mir darüber klar, daß das vorstehende Kapitel keine gültige Deskription bundesdeutscher oder gar internationaler Forschungspraxis darstellt. Dies war auch nicht die Absicht. De facto findet man in psychologischen Fachzeitschriften kaum je etwas anderes als die üblichen klassischen Signifikanztests - meistens ohne jede Reflexion über das Verlässlichkeitsniveau. Man berichtet das Niveau, auf dem die Daten eben „signifikant“ waren, und macht sich schon gar keine Gedanken über den Beta-Fehler. Wozu auch, wenn ohnedies nur „signifikante“ Ergebnisse berichtet werden, und daß dies so ist, dafür sorgt in den meisten Fällen die Annahme- und Ablehnungsstrategie des Herausgebers oder Redakteurs. Und auch das wiederum nicht ohne gewisse Berechtigung, denn mit einer nicht verworfenen Nullhypothese kann man auf

Basis des Neyman-Pearsonschen Systems nichts „beweisen“ - nicht einmal, daß die nicht verworfene Nullhypothese richtig ist, wie oben gezeigt wurde.

So bilden denn die Untersuchungen, in denen die gewünschte statistische Signifikanz nicht erreicht wurde, eine unübersehbare Dunkelziffer, und die Tatsache, daß bei der Publikation von Forschungsergebnissen eine Auswahl getroffen wird - nämlich derart, daß nur die „signifikanten“ erscheinen -, läßt den Verdacht aufkommen, daß der Bestand unseres auf diese Weise „statistisch gesicherten“ Wissens zu einem großen Teil auf Zufallsergebnissen beruht, die eben zufällig die magische Signifikanzschränke überschritten haben. Man kann sich davon leicht einen Eindruck verschaffen, wenn man sich einmal die Mühe macht, eine eigene oder fremde Untersuchung mehrfach in gleicher Weise zu replizieren.

Aufgrund dieses Unbehagens an der gegenwärtigen Forschungspraxis in der Psychologie (und auch anderen Wissenschaften) habe ich versucht, hier nicht nur die Mängel des üblichen klassischen Testverfahrens aufzuzeigen, sondern auch Alternativen hierzu. Direkte Vergleiche der hier vorgestellten Bayes-Verfahren mit der klassischen Test-Statistik geben beispielsweise Edwards, Lindman & Savage (1963), Overall (1969), Slovic & Lichtenstein (1971), Fienberg & Zeller (1975) und Rüppell (1977), wo sich der interessierte Leser näher mit den hier angesprochenen Fragen auseinandersetzen kann; außerdem gibt es auch bereits Statistik-Lehrbücher auf Bayes'scher Basis, wie beispielsweise Winkler (1972), Phillips (1973), Box & Tiao (1973), Novick & Jackson (1974) und DeGroot (1975). Eine gut verständliche (und befürwortende) Einführung in die Bayes-Verfahren gibt Rüppell (1977); sehr kritisch und polemisch setzt sich dagegen Rützel (1979) mit ihnen auseinander. Er bezeichnet das Bayes'sche Hypothesentesten als „privates Hypothesentesten“ und als solches als „unverbindlich und für den öffentlichen wissenschaftlichen Gebrauch abzulehnen“ (S. 224). Er wendet sich vor allem gegen die Verwendung von a-priori-Wahrscheinlichkeiten ( $P(H)$ ) und fordert:

entweder „Neutralität“ im Sinne gleicher a-priori-Wahrscheinlichkeiten (was sich weitgehend mit unserer oben aufgestellten Forderung nach Forschungsökonomie deckt, wenn es auch dort aus anderen Gründen gefordert wurde), oder aber - wenn die Anfangsbestätigungen  $P(H)$  der Hypothesen schon ungleich sind - daß diese unterschiedlichen  $P(H)$  empirisch „nach intersubjektiv eindeutigen Kriterien (sprich: objektiv) festgesetzt werden“ (S. 224). Genau dieser letzten Forderung kommt aber eine besondere Richtung der Bayes-Statistik nach, nämlich die „empirical Bayes methods“, z.B. Maritz (1970).

Auf seiten der Bayesianer zeigen Lindley & Phillips (1976) an einem Beispiel, daß auch der klassische Statistiker nicht ohne ungeprüfte Annahmen auskommt, ja, manche (relativ einfachen) statistischen Probleme wie die Bestim-

mung der Wahrscheinlichkeit des Ausganges der  $(n+1)$ ten Beobachtung aufgrund der Ausgänge der  $n$  vorangegangenen Beobachtungen ohne Zusatzannahmen gar nicht lösen kann.

Kemmnitz (1980) kritisiert die Bayes-Statistik hauptsächlich auf sprachanalytischer Basis, und zwar an dem Konzept der Wahrscheinlichkeit als subjektiver Glaubensgrad. Auf seine Kritik an dem Wahrscheinlichkeitsbegriff der „Subjektivisten“ (die tatsächlich besser von unsicherem „Wissen“ statt von „Glauben“ und „Gefühl“ reden sollten) sind wir bereits bei der Besprechung der Belohnungsfunktionen eingegangen. Hauptkritikpunkt der Gegner der Bayes-Statistik ist die Einbeziehung der a-priori-Wahrscheinlichkeiten, die ihnen - da subjektiv - suspekt und willkürlich erscheint. Das dagegen von den Bayesianern vorgebrachte Argument des Prinzips der Robustheit der Schätzung (principle of stable estimation) erscheint z.B. Rützel (1979) als „Rückzugsgefecht“, das (auch hier weiter vorne zitierte) Beispiel aus Edwards, Lindman & Savage (1963) sowie Lindleys Paradoxon (1957) als künstlich konstruiert und an den Haaren herbeigezogen. Bei der Diskussion dieser Argumentation sollten wir jedoch nicht außer acht lassen, daß hier selbst die Gegner der Bayes-Statistik implizite mit subjektiven a-priori-Wahrscheinlichkeiten operieren, nämlich mit ihrer (subjektiv) geringen Wahrscheinlichkeit für das Auftreten der Bedingungen für Edwards, Lindman & Savages Beispiel oder Lindleys Paradoxon.

über die Kontroverse zwischen „klassischer“ und Bayes-Statistik ist sicherlich das letzte Wort noch nicht gesprochen. Wahrscheinlich (wobei dieser Wahrscheinlichkeitsbegriff ein subjektiver ist!) wird sich ein Kompromiß erreichen lassen, in dem beide Ansätze ihre Existenzberechtigung und ihren Platz finden - vielleicht der klassische (frequentistische) in den Fällen, wo frequentistische Wahrscheinlichkeiten aufgrund hinreichend vieler Beobachtungen möglich und sinnvoll sind, und der subjektive in den Fällen seltener oder einzelner Ereignisse, bei denen zur Beurteilung der Wahrscheinlichkeit auf solche Quellen nicht zurückgegriffen werden kann. Es erschien mir aber sinnvoll und angemessen, in diesem Kapitel auf die neueren Alternativen zu den üblichen klassischen Testverfahren hinzuweisen, zumal dieses Handbuch eher eine Chance haben dürfte, bei anstehenden methodischen Fragen konsultiert zu werden, als ein Zeitschriftenaufsatz.

## Literatur

- Aschenbrenner, K. M. 1977. Komplexes Wahlverhalten: Entscheidungen zwischen multiattributen Alternativen. In: K. D. Hartmann & K. Koeppler (Hg.) Fortschritte der Marktpsychologie, Band 1. Fachbuchhandlung für Psychologie, Frankfurt, S. 21-52.
- Aschenbrenner, K. M. & Wendt, D. 1978. Expectation maximization versus ambition motivation in probability estimation. *Organizational Behavior and Human Performance*, 21, 160-170.
- Bayes, T. 1763, 1958. Essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society*, 53, 370-418. Reprinted in: *Biometrika*, 45, 293-315.
- Blackwell, D. & Dubins, L. 1962. Merging of opinion with increasing information. *Annals of Mathematical Statistics*, 33, 882-886.
- Borcherding, K. 1978. Subjektive Bestimmung der Erträge von Aktien für Entscheidungshilfe bei der Portfolio Selektion: Theoretischer Bezugsrahmen und eine experimentelle Überprüfung. Dissertation, Philosophische Fakultät der Universität Mannheim.
- Box, G. E. P. & Tiao, G. C. 1973. Bayesian inference in statistical analysis. Addison Wesley.
- Bredenkamp, J. 1969. Experiment und Feldexperiment. In: C. F. Graumann (Hg.) *Handbuch der Psychologie in 12 Bänden*, Band 7: Sozialpsychologie, 1. Halbband. Göttingen: Hogrefe.
- Bredenkamp, J. 1970. über Maße der praktischen Signifikanz. *Zeitschrift für Psychologie*, 177, 310-318.
- Bredenkamp, J. 1972. Der Signifikanztest in der psychologischen Forschung. Frankfurt: Akademische Verlagsgesellschaft.
- Carnap, R. & Stegmüller, W. 1959. Induktive Logik und Wahrscheinlichkeit. Wien: Springer.
- Cohen, J. 1977. Statistical power analysis for the behavioral sciences. New York: Academic Press.
- Cook, T. D., Gruder, C. L., Henningan, K. & Flay, B. R. 1979. History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin*, 86, 662-679.
- Coombs, C. H., Raiffa, H. & Thrall, R. M. 1954. Some views on mathematical models and measurement theory. *Psychological Review*, 61, 132-144.
- Dyer, J. S. & Miles, R. F. 1976. An actual application of collective choice theory to the selection of trajectories for the Mariner Jupiter/Saturn 1977 Project. *Operations Research*, 24, 220-244.
- Edwards, W., Guttentag, M. & Snapper, K. 1975. A decision-theoretic approach to evaluation research. In: E. L. Struening & M. Guttentag (Eds) *Handbook of evaluation research*. Vol. 1, Beverly Hills: Sage Publications.

- Edwards, W., Lindman, H. & Savage, L. J. 1963. Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193-242.
- Finetti, B. de. 1937. La prevision: ses lois logiques, ses sources subjectives. *Anuales de l'Institut Henri Poincaré*, Vol. 7, 1964. English Translation: Foresight: Its logical laws, its subjective sources. In: H. E. Kyburg & H. E. Smokler (Eds) *Studies in subjective probability*. New York: Wiley, p. 93-158.
- Finetti, B. de. 1962. Does it make sense to speak of „good probability appraisers“? In: J. Good (Ed.) *The scientist speculates: An anthology of partly-baked ideas*. London: Heinemann.
- Fisher, R. A. 1921. On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society, Series A*, 222, 309-368.
- Fisher, R. A. 1925. *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
- Fisz, M. 1958, 1970. *Wahrscheinlichkeitsrechnung und mathematische Statistik*. Berlin: VEB Verlag der Wissenschaften.
- Groot, M. H. de. 1970. *Optimal statistical decisions*. New York: McGraw-Hill.
- Hagen, K. & Seifert, H. G. 1979. *Methoden der Statistik für Psychologen*. Band 2. Stuttgart: Kohlhammer.
- Hays, W. L. 1973. *Statistics for the social sciences*. New York: Holt, Rinehart & Winston.
- Henning, H. J. & Muthig, K. 1979. *Grundlagen konstruktiver Versuchsplanung - ein Lehrbuch für Psychologen*. München: Kösel.
- Hofstätter, P. R. & Wendt, D. 1966<sup>2</sup>, 1967<sup>3</sup>, 1974<sup>4</sup>. *Quantitative Methoden der Psychologie*. München: Johann Ambrosius Barth.
- Keeney, R. L. & Raiffa, H. 1976. *Decision analysis with multiple conflicting objectives, preferences, and value trade-offs*. New York: John Wiley & Sons.
- Kemmnitz, W. 1980. *Die Funktion von scoring rules in der Bayes-Statistik*. Manuskript. Konstanz: Fachbereich Statistik.
- Kemmnitz, W. 1980. *Kritik des Bayesianismus*. (Archiv für die Gesamte Psychologie). Manuskript. Konstanz: Fachbereich Statistik.
- Kolmogoroff, A. N. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.
- Kutschers, M. F. von. 1972. *Wissenschaftstheorie I, II*. München: Wilhelm Fink.
- Lienert, G. A. 1962<sup>1</sup>, 1973<sup>2</sup>. *Verteilungsfreie Methoden in der Biostatistik*. Meisenheim am Glan: Anton Hain.
- Lindley, D. V. 1957. A statistical paradox. *Biometrika*, 44, 187-192
- Lindley, D. V. & Phillips, L. D. 1976. Inference for a Bernoulli process (a Bayesian view). *The American Statistician*, 30, 112-119.
- Maritz, J. S. 1970. *Empirical Bayes methods*. Methuen's Monographs on Applied Probability and Statistics. London: Methuen.

- Melton, A. W. 1962. Editorial. *Journal of Experimental Psychology*, 64, 553-557.
- Mises, R. von. 1936. *Wahrscheinlichkeit, Statistik und Wahrheit*. Wien: Springer.
- Moroney, M. J. 1951. *Facts from figures*. Penguin Books, 1970. Deutsche Übersetzung: *Einführung in die Statistik*. Buchreihe Orientierung und Entscheidung. München: Oldenbourg.
- Murphy, A. H. & Winkler, R. L. 1970. Scoring rules in probability assessment and evaluation. *Acta Psychologica*, 34, 273-286.
- Murphy, A. H. & Winkler, R. L. 1977. Reliability of subjective probability forecasts of precipitation and temperature. *The Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 26, 4147.
- Nagel, E. 1939. Principles of the theory of probability. *International Encyclopedia of Unified Science*, Vol. I, No. 6. Chicago: University of Chicago Press.
- Neumann, J. von & Morgenstern, O. 1944, 1947<sup>2</sup>. *Theory of games and economic behavior*. Princeton: University Press, 1961. Deutsch: *Spieltheorie und wirtschaftliches Verhalten*. Würzburg: Physika-Verlag.
- Neyman, J. & Pearson, E. S. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 20A, 175 und 263.
- Neyman, J. & Pearson, E. S. 1933. On the testing of statistical hypotheses in relation to probability a priori. *Proceedings of the Cambridge Philosophical Society*, 29, 492.
- Neyman, J. & Pearson, S. E. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A*, 231, 281.
- Novick, M. R. & Jackson, P. H. 1974. *Statistical methods for educational and psychological research*. Manchester: McGraw-Hill.
- Overall, J. E. 1969. Classical statistical hypothesis testing within the context of Bayesian theory. *Psychological Bulletin*, 71, 285-292.
- Phillips, L. D. 1973. *Bayesian statistics for social scientists*. London: Thomas Nelson & Sons Ltd.
- Rüppel, H. 1977. Bayes-Statistik. Eine Alternative zur klassischen Statistik. *Archiv für Psychologie*, 129, 175-186.
- Rützel, E. 1979. Bayes'sches Hypothesentesten und warum die Bayesianer Bias-ianer heißen sollten. *Archiv für Psychologie*, 131, 211-232.
- Savage, L. J. 1954. *The foundations of statistics*. New York: Wiley.
- Schneeweiß, H. 1966. *Entscheidungskriterien bei Risiko*. Berlin: Springer.
- Slovic, P. & Lichtenstein, S. 1971. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6, 649-744.
- Stäel von Holstein, C.-A. S. 1977. The continuous ranked probability score in practice. In: H. Jungermann & G. de Zeeuw (Eds) *Decision making and change in human affairs*. Dordrecht: Reidel.

- Stegmüller, W. 1973. Personelle und statistische Wahrscheinlichkeit. 1. Halbband: Personelle Wahrscheinlichkeit und Rationale Entscheidung. 2. Halbband: Statistisches Schließen - Statistische Begründung - Statistische Analyse. Berlin: Springer-Verlag.
- Wald, A. 1947. Sequential analysis. New York: Wiley.
- Wald, W. 1950. Statistical decision functions. New York: Wiley.
- Weber, Erna. 1964<sup>5</sup>, 1972<sup>7</sup>. Grundriß der biologischen Statistik. Jena: VEB Gustav Fischer.
- Wendt, D. 1970. Utility and risk. *Acta Psychologica*, 34, 214-228.
- Winkler, R. 1972a. Introduction to Bayesian inference and decision. New York: Holt, Rinehart & Winston.
- Winkler, R. L. & Murphy, A. H. 1968. Evaluation of subjective precipitation probability forecasts. In: *Proceedings of the First National Conference on Statistical Meteorology*. Boston: American Meteorological Society.
- Witte, E. 1980. Signifikanztest und statistische Inferenz. Analysen, Probleme, Alternativen. Stuttgart: Enke.



## 5. Kapitel

# Computer-Simulation

*Hans Ueckert*

### 1. Einleitung

In der Geschichte der Psychologie gab es - in einer etwas groben Rasterung - drei chronologische Einschnitte, die die methodologische Entwicklung der Psychologie zu einer selbständigen Einzelwissenschaft kennzeichnen. Der erste war die Einführung des Experiments in die psychologische Forschung durch Wilhelm Wundt vor etwa 100 Jahren, der zweite die Entwicklung der quantitativen Methodik für die Auswertung psychologischer Daten in der ersten Hälfte unseres Jahrhunderts und der dritte die - wenn man so will - „Erfindung“ der Computer-Simulation zur Nachbildung psychischer Vorgänge auf einer elektronischen Rechenanlage vor rund 20 Jahren. Mit jeder dieser methodologischen Entwicklungslinien sind wissenschaftshistorische Begleiterscheinungen verbunden, die als spezifische Herausforderungen des Selbstverständnisses der Psychologie verstanden werden können. Die Experimentalmethodik wurde aus den Naturwissenschaften übernommen und mußte erst dem psychologischen Gegenstandsbereich angepaßt werden; die quantitative Methodik führte in ihrer konsequentesten Weiterentwicklung zu einer mathematischen Psychologie, die letztlich eine wie auch immer zu bewertende „Mathematisierung“ des Menschen - als dem Hauptgegenstand der Psychologie - bedingt; die Computer-Simulation schließlich ist von der Entwicklung einer „künstlichen Intelligenz“ begleitet, die die Maschine nicht nur - wie bisher - auf dem Gebiet materieller Tätigkeiten in Konkurrenz zum Menschen stellt, sondern - zuvor noch ganz unvorstellbar - gerade auch auf dem Gebiet informationeller (oder „geistiger“) Tätigkeiten.

Trotz dieser zunehmenden, durch die empirische Methodenentwicklung bedingten „Entzauberung“ psychischen Geschehens scheint die „technischste“ aller Methoden, die Computer-Simulation, dem Untersuchungsgegenstand der Psychologie näher kommen zu können als die vorangehenden Entwicklungen. Das Experiment ist zwar der primäre Datenlieferant über psychologische Sachverhalte, bleibt aber aufgrund der methodologisch begründeten Künst-

lichkeit seiner Untersuchungsbedingungen oft von jeglicher Alltagswirklichkeit weit entfernt; die Quantifizierbarkeit psychischer Gegebenheiten ist zwar legitimes Interesse wissenschaftlicher Forschung, verfehlt aber ihren Gegenstand so lange, als die „Meßbarkeiten“ des Psychischen ungeklärt oder nicht nachweisbar sind. Die Computer-Simulation dagegen ist eine Methode, die den Untersuchungsgegenstand nicht nur in beliebig fein abstufbarer Weise beschreibbar macht, sondern die die untersuchten Phänomene mit dem Instrument des Rechners vollständig nachzubilden und damit tatsächlich sogar zu reproduzieren gestattet. Dabei spielt es im Grunde keine Rolle, ob die untersuchten psychischen Gegebenheiten quantifizierbar sind oder nicht - sowohl numerisch als auch nicht-numerisch gegebene Daten sind in ihren Wirkungsbedingungen auf dem Rechner reproduzierbar. Es spielt im Prinzip auch keine Rolle, ob die auf einem Rechner erstellte Nachbildung - oder am Beispiel der „künstlichen Intelligenz“ auch Erzeugung - psychischer Vorgänge auf experimentalpsychologisch erhobenen Daten beruht oder aber auf der Intuition oder der Introspektion des Modellkonstruktors (wie beispielsweise für die Entwicklung von Systemen der „künstlichen Intelligenz“ kennzeichnend); beide Arten der Modellbildung sind, wenn auch mit unterschiedlicher „Abbildungstreue“ hinsichtlich ihres Gegenstandes, gleichermaßen durchführbar.

Allgemein betrachtet kann man die Computer-Simulation als eine psychologische Methode definieren, die die Modellierung psychischer Gegebenheiten und/oder Vorgänge auf einem Rechner derart zu realisieren gestattet, daß nicht nur eine beliebig feine Abbildung des Untersuchungsgegenstandes ermöglicht wird, sondern daß dieser darüber hinaus auch in all seinen interessierenden Merkmalen, Eigenschaften und Funktionszusammenhängen nachgebildet werden kann: Das resultierende Modell beschreibt und erklärt nicht nur den gewählten Wirklichkeitsausschnitt, sondern es reproduziert ihn auch in der von der zugrunde gelegten psychologischen Theorie vorhersagbaren Art und Weise.

Mit der Entwicklung der Methode der Computer-Simulation von Anfang an eng verknüpft war eine grundlegend neue psychologische Theorienbildung, deren sichtbare Ausprägung heute die kognitive Psychologie ist. In ihrem Gefolge wurde der Begriff der Information als Grundkategorie psychischen Geschehens eingeführt: Psychisches Geschehen ist nicht das Wahrnehmungsgefüge substanzloser „seelischer“ oder „geistiger“ Vorgänge (wie es die geisteswissenschaftliche Psychologie sehen möchte), ist nicht am Verhalten ablesbares materielles Tun (wie es beispielsweise der Behaviorismus beschreibt), sondern ist ein trotz aller Komplexität aufschlüsselbarer Prozeß der Informationsverarbeitung, deren substanzielle Grundlage anatomisch und physiologisch nachweisbare Strukturen und Funktionen des Nerven- und Sinnessystems sind. Jedoch nicht Neuroanatomie und -physiologie sind der Gegenstandsbe- reich der kognitiven Psychologie, sondern die „höheren“ Prozesse der Auf-

nahme, Verarbeitung (einschließlich Erzeugung) und Ausgabe von Information über diesem materiellen Fundament. Daß einerseits eine mit nicht-physikalischen Eigengesetzlichkeiten operierende Informationsverarbeitung möglich ist, andererseits diese jedoch ohne geeignete materielle Trägersysteme nicht existieren kann, hätte der Mensch in seiner mehrtausendjährigen Beschäftigung mit sich selbst längst erkennen können; offensichtlich bedurfte es dazu erst der Entwicklung des Computers, dessen Fähigkeit zur Informationsverarbeitung geradezu als Existenzbeweis dafür angesehen werden kann, daß nicht-materielle, „geistige“ oder - wie präziser zu sagen ist - „informationelle“ Prozesse mit ihren Eigengesetzlichkeiten in materiellen Systemen realisierbar sind.

Daß dabei der Computer eigentlich gar kein Rechner ist, der mit irgendwelchen numerischen Werten („Zahlen“) operiert, kein System also, das im wörtlichen Sinne „rechnet“, sondern ein System, das Information beliebiger Art verarbeitet (sei diese Information als Zahlen, Buchstaben, Wörter oder was auch immer interpretierbar), diese Erkenntnis bedurfte erst einer intellektuellen Anstrengung zur Beseitigung eines durch die Computerentwicklung und der anfänglichen Rechnerverwendung bedingten Vorurteils. Der Computer ist nichts anderes als eine Maschine, die aufgrund ihrer jeweiligen Programmierung Zeichen in für sie „lesbarer“ Form verarbeitet, gleichgültig, was immer diese Zeichen „bezeichnen“ mögen.

Wenn die Denkpsychologie beispielsweise den Prozeß des menschlichen Problemlösens als einen Prozeß der sukzessiven Verarbeitung von Information beschreiben und erklären kann, dann ist es ein naheliegender Schritt, diesen Prozeß auch auf einer Maschine nachzubilden, die zur Informationsverarbeitung aufgrund ihrer Konstruktionsmerkmale und ihrer diesbezüglichen Programmierbarkeit fähig ist. Genau dieser Schritt stand auch am Anfang der Entwicklung der Computer-Simulation als psychologische Methode, wie die ersten Arbeiten von Simon, dem eigentlichen Begründer der modernen kognitiven Psychologie, und seinen Mitarbeitern zeigen (vgl. beispielsweise Newell, Shaw & Simon, 1958).

In ihrer weiteren Entwicklung hat die Computer-Simulation - und in ihrem Gefolge auch die „künstliche Intelligenz“ - eine Vielzahl von Modellen psychischer Vorgänge geliefert, die in ihrer Gesamtheit einerseits eine Herausforderung an den Menschen bezüglich seiner „intellektuellen Einzigartigkeit“ darstellen, andererseits aber ohne eine genauere Kenntnis der Grundlagen, Techniken und Anwendungsmöglichkeiten der Computer-Simulation als wissenschaftliche Methode nicht angemessen beurteilt werden können.

## 2. Das Paradigma der Computer-Simulation in der Psychologie

Der vielfältige Gebrauch der Computer-Simulation in den verschiedensten Wissenschafts- und Anwendungsbereichen hat zur Folge, daß man nicht von dem Paradigma der Computer-Simulation sprechen kann, sondern nur von unterschiedlichen Paradigmen entsprechend ihrer Verwendungsweisen in den einzelnen Gebieten.

### 2.1 Zur Klassifikation von Simulationsmodellen

Eine systematische Klassifikation von Simulationsmodellen sollte sowohl von formalen als auch von inhaltlichen Kriterien ausgehen können. Bisher vorgelegte Klassifikationen betonen entweder den einen oder den anderen Ausgangspunkt und variieren daher von Autor zu Autor.

Nach *formalen* methodologischen Kriterien unterscheidet beispielsweise Harbordt (1974, S. 22-29)

- (1) statische und dynamische,
- (2) deterministische und indeterministische,
- (3) quantitative und qualitative,
- (4) analytische und synthetische
- (5) Erkundungs- und Entscheidungsmodelle.

(1) *Statische* Modelle sind in ihren Abbildungseigenschaften auf einen Zeitpunkt bezogen, während der Zeitablauf über mehreren Zeitpunkten *dynamische* Modelle kennzeichnet. Beispielsweise ist die funktionale Verknüpfung  $f$  der Variablen  $X_1$ ,  $X_2$  und  $Y$  zum Zeitpunkt  $t$  mit der Gleichung

$$Y(t) = f[X_1(t), X_2(t)]$$

ein statisches Modell, bei dem die Zeitvariable auch weggelassen werden kann, während die Gleichung

$$Y(t) = f[X_1(t), X_2(t-1)]$$

ein dynamisches Modell bezeichnet, bei dem der Wert der Variable  $Y$  zum Zeitpunkt  $t$  von der funktionalen Beziehung zwischen den Werten der Variable  $X_1$  zum gleichen Zeitpunkt  $t$  und der Variable  $X_2$  zum vorangehenden Zeitpunkt  $t-1$  abhängt. In aller Regel sind Simulationsmodelle als derartige dynamische Modelle konzipiert.

(2) *Deterministische* Modelle sind in ihrem Eingabe-Ausgabe-Verhalten eindeutig bestimmt, d.h. bei bestimmten Eingabewerten einer Variable  $X$  sind die Ausgabewerte der Variable  $Y$  exakt vorhersagbar. Bei *indeterministischen* Mo-

dellen ist diese Vorhersagbarkeit nicht mehr eindeutig gegeben. Nach Harbordt (1974, S. 23-25) kann man hier noch zwischen stochastischen und probabilistischen Modellen unterscheiden. *Stochastische* Modelle enthalten eine oder mehrere Zufallsvariable, deren Werte nur mit bestimmten Auftretens-Wahrscheinlichkeiten vorhersagbar sind, wie z.B. in der Gleichung

$$Y = f(X, Z),$$

in der das Ausgabeverhalten der Variable Y von der funktionalen Beziehung zwischen der (deterministischen) Variable X und der (stochastischen) Zufallsvariable Z abhängt.

*Probabilistische* Modelle sind demgegenüber dadurch gekennzeichnet, daß die Ausgabevariable Y selbst eine Zufallsvariable ist, deren Wahrscheinlichkeitsverteilung P(Y) aufgrund ihrer funktionalen Beziehung zu deterministischen Variablen  $X_i$  bekannt ist. Solche Modelle werden auch als „Monte-Carlo-Simulationen“ bezeichnet, da die tatsächliche Modellausgabe Y aufgrund eines gleichverteilten Zufallsgenerators unter Berücksichtigung des funktionalen Zusammenhangs der Eingabevariablen  $X_i$  erzeugt wird. - In der Mehrzahl der Fälle sind Simulationsmodelle in der Psychologie deterministische Modelle, ggf. um gewisse stochastische oder probabilistische Komponenten ergänzt.

(3) *Quantitative* Modelle verwenden in ihren Abbildungseigenschaften, insbesondere für das Ausgabeverhalten, numerische Variablen (meist mit mindestens Intervallskalenniveau), während *qualitative* Modelle die funktionalen Zusammenhänge zwischen nicht-numerischen Variablen nachbilden (numerisch als Nominal- und/oder Ordinalskalenniveau beschreibbar). Wie im weiteren gezeigt werden soll, bedient sich die Simulationsmethode in der Psychologie in bevorzugter Weise der qualitativen, nicht-numerischen Modellierung ihrer Untersuchungsgegenstände.

(4) Die Unterscheidung zwischen analytischen und synthetischen Modellen der Computer-Simulation ist - wie auch die folgende zwischen Erkundungs- und Entscheidungsmodellen - in gewisser Weise willkürlich. Der Unterschied liegt eher in den Vorgehensweisen des Modellkonstruktors als in den Modelleigenschaften selbst.

*Analytische* Modelle gehen von dem beobachtbaren Gesamtverhalten des zu modellierenden Systems aus und versuchen dieses auf das Zusammenwirken seiner mehr oder minder gut beobachtbaren Komponenten zurückzuführen. *Synthetische* Modelle können dagegen von einer ausreichenden Kenntnis solcher Komponenten ausgehen und das aus diesen zusammensetzbare Gesamtverhalten zu reproduzieren versuchen. - Charakteristisch für die psychologische Simulationsforschung dürfte aufgrund des noch bescheidenen Wissensstandes der Psychologie die analytische Vorgehensweise sein, auch wenn die

synthetische Modellierbarkeit als erstrebenswertes wissenschaftliches Ziel anzusehen sein mag.

(5) Praktisch gesehen fällt die Unterscheidung zwischen Erkundungs- und Entscheidungsmodellen mit der zwischen analytischen und synthetischen Simulationsmodellen zusammen.

*Erkundungsmodelle* wollen einen noch nicht hinreichend untersuchten Gegenstandsbereich mit Hilfe der Simulationsmethode zugänglich machen, in der Regel also ausgehend vom beobachtbaren Gesamtverhalten und dessen Zurückführbarkeit auf seine Komponenten (analytische Modellierung). *Entscheidungsmodelle* sollen dagegen die Möglichkeit eröffnen, zwischen Modellvarianten zu entscheiden (und ggf. zu wählen), deren Verhalten sich in eindeutiger Weise aus bestimmten Komponenten erzeugen läßt (synthetische Modellierung), meist mit der Maßgabe, eine optimale Modellvariante herauszufinden (wie beispielsweise in der Operationsforschung). - Aus den gleichen Gründen wie zuvor herrschen in der Psychologie Erkundungsmodelle vor, obgleich es gerade für die Anwendbarkeit der Simulationsmethode in der psychologischen Praxis - beispielsweise in pädagogischen und in therapeutischen Bereichen - wünschenswert sein dürfte, vermehrt auch mit Entscheidungsmodellen arbeiten zu können.

Eine andere, nicht nach formalen, sondern nach *inhaltlichen* Kriterien ausgerichtete Klassifikation von Simulationsmodellen haben Dutton & Starbuck (1971) entwickelt, die für ihre Verwendbarkeit besonders in den Sozialwissenschaften spricht. Dutton & Starbuck unterscheiden nach den Gegenständen - oder „Merkmalsträgern“ - und deren Zusammenhängen, die Bezugs- und Anwendungsbereich zugleich für die Entwicklung von Simulationsmodellen sind. Danach sind Gegenstands- und Merkmalsbereich der „Computer-Simulation menschlichen Verhaltens“ (so der Titel des Sammelbandes von Dutton & Starbuck):

- (1) Individuen (im Sinne von Einzelfalluntersuchungen),
- (2) interagierende Individuen,
- (3) aggregierte Individuen (im Sinne von in Gruppen zusammengefaßten Individuen),
- (4) aggregierte und interagierende Individuen.

Aus dieser Übersicht ist zu entnehmen, daß mit den vier Kategorien der gesamte Bereich der Sozialwissenschaften - von der Psychologie über Soziologie und Politologie bis hin zu den Wirtschaftswissenschaften - abgedeckt werden kann. Die Psychologie ist verständlicherweise mehr in den Kategorien (1) und (2), Individuen und interagierende Individuen vertreten, wie die Beiträge in dem Sammelband von Dutton & Starbuck zeigen, aber auch, wie sich aus der von den Autoren ausgearbeiteten Bibliographie von über 2000 Titeln

der bis zum Ende der 60er Jahre erschienenen Arbeiten zur Computer-Simulation ablesen läßt.

Zusammenfassend kann man sagen, daß die Computer-Simulation in der Psychologie den Typ der dynamischen, deterministischen, qualitativen und analytischen Erkundungsmodelle zur Nachbildung menschlichen Verhaltens am Beispiel von einzelnen und interagierenden Individuen bevorzugt. An dem Programmbeispiel des nächsten Abschnitts soll diese Typologie, aus der sich das Paradigma der Computer-Simulation in seiner in der Psychologie vorherrschenden Form ableiten läßt, illustrativ konkretisiert werden.

## 2.2 Programmbeispiel: „Simple Concept Attainment“

Ein bevorzugtes Gebiet mathematischer Modellbildungen in der Psychologie ist die experimentelle Begriffsbildungsforschung. In einer illustrativen Arbeit haben Gregg & Simon (1967) am Beispiel der einfachen Begriffsbildung („Simple Concept Attainment“) eine Gegenüberstellung von Prozeßmodellen und stochastischen Theorien vorgenommen, die sich für eine Einführung in das Paradigma der Computer-Simulation in besonderer Weise eignet, da sich hieraus sowohl die Methodik als auch die Problematik des Verfahrens veranschaulichen lassen.

Im experimentellen Design zur einfachen Begriffsbildung (genauer: „Begriffsfindung“) werden der Versuchsperson  $n$ -dimensionale Reize mit je zwei möglichen Ausprägungen vorgegeben (z.B. „großer roter Kreis“ mit den drei Dimensionen „Größe“, „Farbe“ und „Form“ und deren zweifachen Ausprägungsmöglichkeiten „klein, groß“, „rot, blau“ und „Kreis, Quadrat“). Relevant für die Begriffsbildung ist genau eine Ausprägung einer bestimmten Dimension (z.B. „blau“), die die Versuchsperson im Verlauf des Experiments herauszufinden hat, indem sie mit „ja“ oder „nein“ antwortet, je nachdem, ob das von ihr vermutete Konzept in dem vom Versuchsleiter vorgelegten Beispiel enthalten ist oder nicht. Auf jede Antwort der Versuchsperson gibt der Versuchsleiter eine Rückmeldung darüber, ob die Antwort „richtig“ oder „falsch“ ist. Der Versuch wird so lange fortgesetzt, bis die Anzahl richtiger Antworten der Versuchsperson hintereinander einen bestimmten, vom Versuchsleiter festgelegten Kriteriumswert erreicht hat.

### 2.2.1 Flußdiagrammdarstellung

Ein nützlicher vorbereitender Schritt in der Entwicklung eines Simulationsmodells ist - noch vor der eigentlichen Modelldarstellung in einer bestimmten Programmiersprache - die Ausarbeitung einer Verlaufsskizze des Modellver-

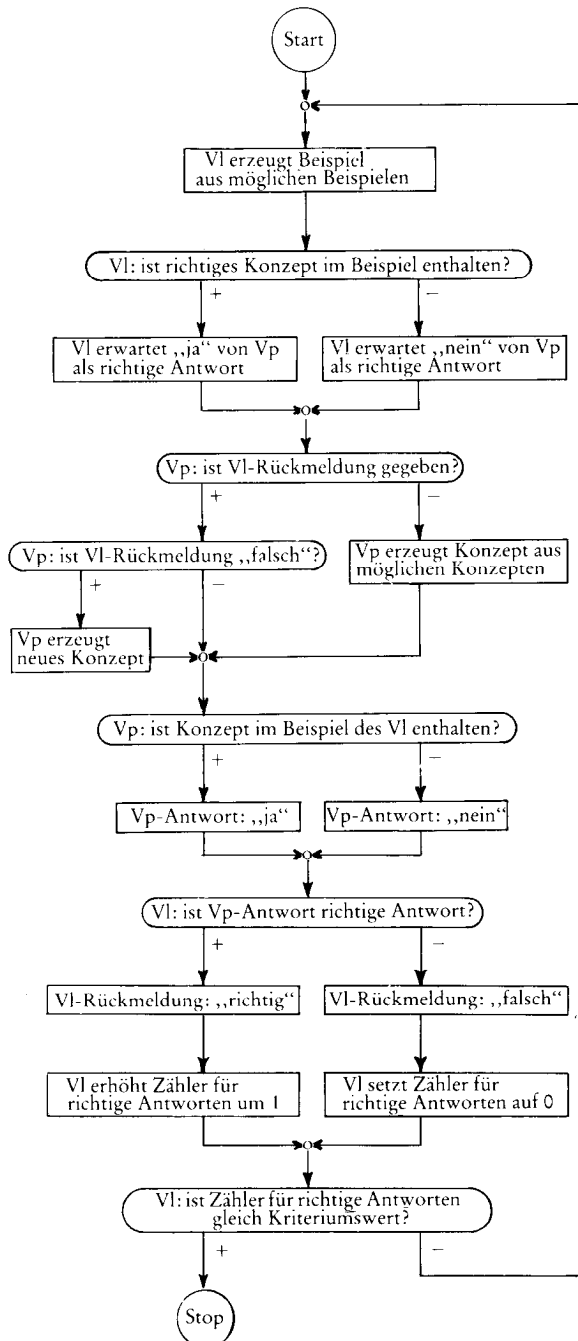


Abb. 1: Flußdiagramm zur einfachen Begriffsbildung.



haltens in Form eines Flußdiagramms. Ein *Flußdiagramm* ist die graphische Darstellung des Verhaltensablaufs aufgrund der im Modell auszuführenden Operationen und Entscheidungen. Operationen werden in einem rechteckigen Kästchen mit je einem Eingang und Ausgang, *Entscheidungen* über alternative Wege in einem ovalen Kästchen mit einem Eingang und zwei Ausgängen dargestellt; Richtung und Reihenfolge der Operationen und Entscheidungen wird durch Pfeile zwischen den Kästchen angedeutet.

In Abb. 1 ist nun in Form eines Flußdiagramms der Versuchsablauf zur einfachen Begriffsbildung schon recht detailliert dargestellt. Klar erkennbar sind drei Hauptphasen des Experiments: (1) die Aktivitäten des Versuchsleiters (VI) zur Vorgabe eines Reizbeispiels, (2) die Handlungen der Versuchsperson (Vp) für ihre Antwort auf die Beispielvorgabe und (3) die Operationen und Entscheidungen des Versuchsleiters für eine Rückmeldung an die Versuchsperson und zur eventuellen Beendigung des Experiments.

Die Entwicklung eines angemessenen Simulationsmodells für den interessierenden Gegenstandsbereich kann jedoch nicht bei einer Flußdiagrammdarstellung stehen bleiben, da sie - trotz aller Ausführlichkeit - noch zu ungenau ist. Was beispielsweise „VI erzeugt Beispiel aus möglichen Beispielen“ oder „Vp erzeugt Konzept aus möglichen Konzepten“ genau heißen soll, d.h. welche konkreten Operationen im einzelnen hier wirklich auszuführen sind, ist noch viel zu unbestimmt, um schon von einem Simulationsmodell der einfachen Begriffsbildung sprechen zu können, das das individuelle Verhalten von Versuchsleiter und Versuchsperson in allen interessierenden Aspekten nachzubilden gestattet.

Der wichtigste Schritt in der Entwicklung eines Simulationsmodells ist daher die Programmierung des Modells in einer geeigneten Programmiersprache, um den Verhaltensablauf tatsächlich auf einem Rechner - und zwar Schritt für Schritt - beobachten zu können. In den folgenden Unterabschnitten wird eine derartige Programmierung exemplarisch für das „Simple Concept Attainment“ - in Anlehnung an die Arbeit von Gregg & Simon (1967) - vorgeführt, wobei die Programmiersprache LOGO ihrer Einfachheit und leichten Verständlichkeit wegen als Illustrationssprache verwendet werden soll.

### 2.2.2 Das Hauptprogramm (Versuchsablaufprogramm)

In Abb. 2 ist das Hauptprogramm des „Simple Concept Attainment“ wiedergegeben, wie es sich in der Programmiersprache LOGO formulieren läßt; es beschreibt den Versuchsablauf in seinen drei Hauptphasen, wie sie im vorigen Unterabschnitt herausgestellt wurden.

```

to SIMPLE.CONCEPT.ATTAINMENT
:TRUE.CONCEPT:
:POSSIBLE.CONCEPTS:
:CRITERION.VALUE:
:MODEL.VARIANT:
10 make „FEEDBACK“ :empty:
20 make „RIGHT.ANSWER.COUNT“ 0
30 print :empty:
40 print sentence „INSTANCE IS A“ EXPERIMENTER'S.INSTANCE
50 print sentence „*** I SAY“ SUBJECT'S.ANSWER
60 print sentence „YOUR ANSWER IS“ EXPERIMENTER'S.FEEDBACK
70 if :RIGHT.ANSWER.COUNT: = :CRITERION.VALUE:
    then stop
    else go to line 30
end

```

Abb. 2: Hauptprogramm (Versuchsablaufprogramm) zur einfachen Begriffsbildung.

Bevor auf inhaltliche Einzelheiten dieses Programms eingegangen werden kann, seien einige Erläuterungen zur LOGO-Programmierung vorangestellt:

(1) Alle hier *klein* geschriebenen Ausdrücke und die Zahlen, die Rechenzeichen (+, -, \*, /), das Gleichheitszeichen (=) sowie die numerischen Prädikate (<, >) gehören zum Grundvokabular von LOGO. Die meisten dieser Ausdrücke sind schon aufgrund ihrer Wortwahl selbsterklärend, so daß auf ihre genaue Beschreibung verzichtet werden kann. (Anmerkung: In LOGO selbst gibt es die Unterscheidung zwischen Groß- und Kleinschreibung nicht in dieser, sondern in der umgekehrten Form, d.h. die zum Grundvokabular von LOGO gehörenden Ausdrücke sind stets groß zu schreiben, während die benutzerdefinierten Ausdrücke wahlweise groß oder klein geschrieben werden können. Die hier gewählte Schreibweise dient lediglich zu Darstellungszwecken für die vorliegende Arbeit.)

(2) Die *groß* geschriebenen Ausdrücke sind vom Benutzer für seine Programmierung frei wählbare Bezeichnungen, für die es folgende syntaktische Vereinbarungen gibt:

(a) Ein in Anführungszeichen eingeschlossener Ausdruck wird von LOGO als eine wörtliche Zeichenfolge - als ein „Literele“ - gelesen und nicht weiter ausgewertet. Literale sind - als „LOGO-Wörter“ oder als „LOGO-Sätze“, die aus LOGO-Wörtern bestehen - die Grunddaten bzw. die Informationen, die von LOGO verarbeitet werden können; dabei werden die Zahlen als LOGO-Wörter dargestellt, ohne daß sie in Anführungszeichen zu setzen sind. Der Ausdruck „FEEDBACK“ in Programmzeile 10 von Abb. 2 ist ein Beispiel für ein LOGO-Wort, der Ausdruck „INSTANCE IS

A“ in Programmzeile 40 ein Beispiel für einen LOGO-Satz, während die Zahl 0 in Programmzeile 20 ein Beispiel für eine ohne Anführungsstriche geschriebene LOGO-Zahl ist (wie übrigens auch alle Zeilennummern).

(b) Ein in Doppelpunkte eingeschlossener Ausdruck wird von LOGO als der Name einer Variable interpretiert, deren *Werte* irgendwelche LOGO-Wörter oder LOGO-Sätze sein können (einschließlich des leeren Ausdrucks „“, für den auch der LOGO-Name :empty: stehen kann). Beispielsweise ist der Ausdruck :RIGHT.ANSWER.COUNT: in Programmzeile 70 der Name einer Variable, deren Wert in Programmzeile 20 mit der LOGO-Anweisung ‚make‘ und dem entsprechenden LOGO-Wort „RIGHT.ANSWER.COUNT“ zunächst einmal auf 0 gesetzt wird.

(c) Alle übrigen Ausdrücke werden von LOGO als auszuführende *Funktionen* („Prozeduren“) verstanden, die entweder eine Anweisung darstellen (Beispiel: ‚print‘ in den Programmzeilen 30-60) oder eine Operation mit LOGO-Wörtern oder LOGO-Sätzen bilden (Beispiel: ‚sentence‘ in den Programmzeilen 40-60). Der Unterschied zwischen Anweisungen und Operationen wird aus den weiteren Beispielen noch ersichtlich werden. (Anmerkung: Die LOGO-Ausdrücke ‚of‘ und ‚and‘ - sie kommen in den weiteren Beispielen wiederholt vor - sind keine LOGO-Funktionen, sondern Füllwörter [„noise words“], die von LOGO überlesen werden, zur besseren Lesbarkeit von LOGO-Programmen jedoch beitragen können.)

(3) Die vom Benutzer für sein Programm zu definierenden Funktionen werden mit der LOGO-Anweisung ‚to‘ eröffnet und mit der LOGO-Anweisung ‚end‘ abgeschlossen. Dazwischen stehen die mit einer fortlaufenden, frei wählbaren Zeilennummerierung versehenen Anweisungen und Operationen, die innerhalb der Funktion zum Zeitpunkt ihres Aufrufs ausgeführt werden sollen. Der vom Benutzer für seine Funktionsdefinition zu vereinbarende Name wird anschließend an die LOGO-Anweisung ‚to‘ geschrieben, gefolgt von einer Angabe der in Doppelpunkte eingeschlossenen Variablennamen, die für den späteren Funktionsaufruf von Interesse sind. (Anmerkung: Will man - wie im Beispiel mit dem Wort SIMPLE.CONCEPT.ATTAINMENT - längere Funktionsnamen bilden, so sind die einzelnen Wörter am besten mit einem Punkt zu einem LOGO-Wort aneinanderzuhängen, da der sonst übliche Bindestrich in LOGO schon als Minuszeichen vergeben ist; gleiches gilt auch für die Bildung längerer Variablennamen.)

Nach diesen Erläuterungen dürfte das LOGO-Programm von Abb. 2 schon leichter zu verstehen sein: Das Programm hat den Funktionsnamen SIMPLE.CONCEPT.ATTAINMENT mit den Variablen

:TRUE.CONCEPT: (das vom Versuchsleiter gewählte und von der Versuchsperson zu findende Konzept),

:POSSIBLE.CONCEPTS: (die vorzugebenden  $2n$  Werte für die  $n$ -dimensionalen Reizbeispiele),  
 :CRITERION.VALUE: (der für die Beendigung des Versuchs vom Versuchsleiter gewählte Kriteriumswert),  
 :MODEL.VARIANT: (die für das Versuchspersonenverhalten formulierbaren Modellvarianten; Einzelheiten dazu weiter unten).

In den Programmzeilen 10 und 20 werden die Voreinstellungen für den Versuchsbeginn vorgenommen, d.h. die Rückmeldung seitens des Versuchsleiters ist zunächst leer und die Anzahl richtiger Antworten der Versuchsperson 0. Kernstück des Programms sind die Druckanweisungen der Programmzeilen 30-60 (sie entsprechen den in der Flußdiagrammdarstellung genannten drei Hauptphasen des Experiments): Zunächst wird das Versuchsleiterbeispiel in Form des LOGO-Satzes „INSTANCE IS A . . .“ ausgegeben, dann folgt die

SIMPLE.CONCEPT.ATTAINMENT

„BLUE“

„SMALL BIG RED BLUE CIRCLE SQUARE“

5

„GLOBAL.CONSISTENCY.MODEL“

INSTANCE IS A BIG BLUE CIRCLE

\*\*\*\* I SAY NO

YOUR ANSWER IS WRONG

INSTANCE IS A SMALL RED SQUARE

\*\*\*\* I SAY NO

YOUR ANSWER IS RIGHT

INSTANCE IS A SMALL RED CIRCLE

\*\*\*\* I SAY NO

YOUR ANSWER IS RIGHT

INSTANCE IS A BIG BLUE SQUARE

\*\*\*\* I SAY YES

YOUR ANSWER IS RIGHT

INSTANCE IS A BIG RED SQUARE

\*\*\*\* I SAY NO

YOUR ANSWER IS RIGHT

INSTANCE IS A SMALL BLUE SQUARE

\*\*\*\* I SAY YES

YOUR ANSWER IS RIGHT

Abb. 3: Probelauf des SIMPLE.CONCEPT.ATTAINMENT-Programms.

Antwort der Versuchsperson in Form von „\*\*\*\* I SAY . . .“ und anschließend gibt der Versuchsleiter seine Rückmeldung in Form von „YOUR ANSWER IS...“. Programmzeile 70 dient als Test dafür, ob der Versuch durch Kriteriumserreichung abgeschlossen werden kann (,stop‘) oder noch fortzusetzen ist (,go to line 30‘); dies entspricht der letzten Zeile in dem Flußdiagramm von Abb. 1.

Ein Versuchsablauf könnte beispielsweise das in Abb. 3 gezeigte Aussehen haben.

Dieses Beispiel liest sich durchaus wie ein tatsächliches Versuchsablaufprotokoll einer Experimentalsitzung und hat zumindest aus dieser - noch oberflächlichen - Sicht Anspruch auf eine realitätsgerechte Beschreibung beobachtbaren Verhaltens.

### 2.2.3 Zur „Binnenstruktur“ der Informationsverarbeitung

Interessanter für die Computer-Simulation kognitiver Vorgänge ist stets die „Binnenstruktur“ der innerhalb der Versuchsperson ablaufenden Informationsverarbeitung. Tatsächlich würde ja auch das in Abb. 2 formulierte Programm noch nicht laufen, wenn man es so wie im Kopf von Abb. 3 angegeben in LOGO aufrufe, denn in den Programmzeilen 40-60 des Hauptprogramms sind die drei Funktionen EXPERIMENTER'S.INSTANCE, SUBJECT'S.ANSWER und EXPERIMENTER'S.FEEDBACK noch vom Benutzer zu definieren, bevor er das Programm starten kann.

In den nachfolgenden Abb. 4-6 sind diese „Unterprogramme“ des SIMPLE.CONCEPT.ATTAINMENT wiedergegeben. Sie sind die eigentlichen „Simulationsprogramme“ für das Versuchsleiterverhalten einerseits und das Versuchspersonenverhalten andererseits.

```

to EXPERIMENTER'S.INSTANCE
10  make „INSTANCE“
      GENERATE.INSTANCE from :POSSIBLE.CONCEPTS:
20  if CONTAINS :INSTANCE: :TRUE.CONCEPT:
      then make „RIGHT.ANSWER“ „YES“
      else make „RIGHT.ANSWER“ „NO“
30  output :INSTANCE:
end

```

Abb. 4: Unterprogramm für das Erzeugen eines Beispiels durch den Versuchsleiter.

Wie ein Vergleich zeigt, entspricht EXPERIMENTER'S.INSTANCE (Abb. 4) dem oberen, dreizeiligen Teil des Flußdiagramms in Abb. 1. In Programmzeile 10 von EXPERIMENTER'S.INSTANCE wird das der Versuchsperson vorzulegende Beispiel mit dem - vom Benutzer noch zu definierenden - Teilprogramm GENERATE.INSTANCE erzeugt. In Programmzeile 20 wird geprüft, ob das zu lernende „wirkliche Konzept“ in diesem Beispiel enthalten ist oder nicht, worauf der Versuchsleiter als „richtige Antwort“ „ja“ oder „nein“ von der Versuchsperson erwarten wird, wenn diese das Konzept gefunden hat. In Programmzeile 30 schließlich wird das Beispiel ausgegeben, um vom Hauptprogramm in Programmzeile 40 (Abb. 2) in Form des LOGO-Satzes „INSTANCE IS A :INSTANCE:“ ausgedruckt zu werden (vgl. auch Abb. 3).

```

to SUBJECT'S.ANSWER
10 if not :FEEDBACK: = :empty:
    then go to line 40
20 make „MY.POSSIBLE.CONCEPTS“ :POSSIBLE.CONCEPTS:
30 make „MY.CONCEPT“
    GENERATE.CONCEPT from :MY.POSSIBLE.CONCEPTS:
40 if :FEEDBACK: = „WRONG“
    then do :MODEL.VARIANT:
50 if CONTAINS :INSTANCE: :MY.CONCEPT:
    then make „ANSWER“ „YES“
    else make „ANSWER“ „NO“
60 make „OLD.INSTANCE“ :INSTANCE:
70 output :ANSWER:
end

```

Abb. 5: Unterprogramm für die Antwort der Versuchsperson.

SUBJECT'S.ANSWER (Abb. 5), das dem mittleren, fünfzeiligen Teil des Flußdiagramms in Abb. 1 entspricht, ist dasjenige Unterprogramm, das von zentralem Interesse für die in einer Versuchsperson bei der einfachen Begriffsbildung ablaufenden Prozesse der Informationsverarbeitung ist. Zunächst wird die Versuchsperson, da sie im ersten Versuchsdurchgang noch kein :FEEDBACK: vom Versuchsleiter bekommen hat (vgl. Programmzeile 10 von SUBJECT'S.ANSWER), ihr mögliches Konzeptrepertoire aus den vom Versuchsleiter vorgegebenen möglichen Konzepten bilden (Programmzeile 20). Dann wählt sie sich ein eigenes Konzept mit dem - noch zu definierenden - Teilprogramm GENERATE.CONCEPT aus ihren möglichen Konzepten aus (Programmzeile 30). Da im ersten Versuchsdurchgang aufgrund des noch fehlenden :FEEDBACK: die Programmzeile 40 nicht zur Ausführung kommt, wird die Versuchsperson laut Programmzeile 50 ihre Antwort generieren, und

zwar derart, daß sie „ja“ sagen wird, wenn ihr Konzept in dem vorgelegten Beispiel enthalten ist, und „nein“, wenn dies nicht der Fall ist. In Programmzeile 70 gibt sie diese Antwort dann aus, und zwar in Form des LOGO-Satzes „\*\*\*\* I SAY :ANSWER:“ aufgrund der Programmzeile 50 des Hauptprogramms (vgl. Abb. 2 und für den Probelauf Abb. 3). Programmzeile 60 von SUBJECT'S.ANSWER dient lediglich dem späteren Erinnern des gerade vorgelegten Beispiels und ist eigentlich nur für eine bestimmte Modellvariante des Versuchspersonenverhaltens relevant.

Hat die Versuchsperson ihre Antwort gegeben, dann verlangt die Experimentalanordnung ein :FEEDBACK: auf diese Antwort durch den Versuchsleiter (vgl. Programmzeile 60 in SIMPLE.CONCEPT.ATTAINMENT, Abb. 2). Das Unterprogramm EXPERIMENTER'S.FEEDBACK (Abb. 6), das dem unteren, dreizeiligen Teil des Flußdiagramms in Abb. 1 entspricht (außer dessen letzter Zeile), erfüllt diesen Zweck.

```

to EXPERIMENTER'S.FEEDBACK
10 if :ANSWER: = :RIGHT.ANSWER:
    then make „FEEDBACK“ „RIGHT“
    else make „FEEDBACK“ „WRONG“
20 if :FEEDBACK: = „RIGHT“
    then make „RIGHT.ANSWER.COUNT“:RIGHT.ANSWER.COUNT: + 1
    else make „RIGHT.ANSWER.COUNT“ 0
30 output :FEEDBACK:
end

```

Abb. 6: Unterprogramm für die Rückmeldung über die Versuchspersonenantwort.

In Programmzeile 10 wird geprüft, ob die Antwort der Versuchsperson der zu erwartenden richtigen Antwort entspricht, wenn die Versuchsperson das zu suchende Konzept gefunden haben sollte; im positiven Fall wird die Rückmeldung auf „richtig“ gesetzt, im negativen Fall auf „falsch“, und in Programmzeile 30 wird dann der entsprechende Wert ausgegeben. Zuvor jedoch (Programmzeile 20) wird noch der Zähler für die richtigen Antworten der Versuchsperson um 1 erhöht, wenn das :FEEDBACK: „richtig“ ergab, bzw. auf 0 gesetzt, wenn es „falsch“ ergab. Dieser Zähler wird für die Beendigung des Experiments benötigt (vgl. Programmzeile 70 des Hauptprogramms, Abb. 2).

#### 2.2.4 Die Modellvarianten

Die Modellvarianten, die in den einzelnen Versuchsdurchgängen immer dann zur Ausführung kommen, wenn das :FEEDBACK: durch den Versuchsleiter

„falsch“ ist (Programmzeile 40 von SUBJECT'S.ANSWER), sind das Kernstück der Theorie, die man zur psychologischen Erklärung der einfachen Begriffsbildung heranziehen kann (sie beinhalten einen Aspekt des Simulationsmodells, der in der Flußdiagrammdarstellung von Abb. 1 nur mit einem kleinen, unscheinbaren Operationskästchen namens „Vp erzeugt neues Konzept“ ausgewiesen ist). Gregg & Simon (1967) formulierten in ihrer Arbeit vier solcher Modellvarianten (und das sind bei weitem nicht alle, die man sich für die einfache Begriffsbildung ausdenken kann), deren Programmierung in LOGO in den folgenden Abb. 7-10 wiedergegeben ist.

```
to GLOBAL.CONSISTENCY.MODEL
10 make „MY.POSSIBLE.CONCEPTS“
    REMOVE :MY.CONCEPT: from :MY.POSSIBLE.CONCEPTS:
20 make „NEW.CONCEPT“
    GENERATE.CONCEPT from :MY.POSSIBLE.CONCEPTS:
30 make „MY.CONCEPT“ :NEW.CONCEPT:
end
```

Abb. 7: Erste Modellvariante der einfachen Begriffsbildung.

Das Modell für ein möglichst optimales Versuchspersonenverhalten ist das GLOBAL.CONSISTENCY.MODEL (Abb. 7): Die Versuchsperson entfernt ihr augenblickliches Konzept aus ihrem möglichen Konzeptrepertoire (Programmzeile 10), wenn das :FEEDBACK: durch den Versuchsleiter „falsch“ war (vgl. Programmzeile 40 von SUBJECT'S.ANSWER, Abb. 5, wo die jeweilige Modellvariante durch die LOGO-Anweisung ‚do‘ aufgerufen wird). Dann bildet die Versuchsperson ein neues Konzept aus den verbleibenden möglichen Konzepten (Programmzeile 20 von Abb. 7) und macht dieses zu ihrem Konzept (Programmzeile 30), mit dem sie weiterarbeiten wird (vgl. Programmzeile 50 von SUBJECT'S.ANSWER, Abb. 5). Es ist klar, daß mit dieser Strategie das gesuchte Konzept in jedem Fall von der Versuchsperson gefunden werden kann, da mit der Zeit alle falschen Konzepte aus der Menge der möglichen Konzepte ausgeschlossen werden.

```
to LOCAL.CONSISTENCY.MODEL
20 make „NEW.CONCEPT“
    GENERATE.CONCEPT from :MY.POSSIBLE.CONCEPTS:
30 if CONTAINS :OLD.INSTANCE: :NEW.CONCEPT:
    then go to line 20
    else make „MY.CONCEPT“ :NEW.CONCEPT:
end
```

Abb. 8: Zweite Modellvariante der einfachen Begriffsbildung.



Etwas weniger optimal ist die Strategie des LOCAL.CONSISTENCY.MODEL (Abb. 8). Hier wird lediglich geprüft, nachdem die Versuchsperson ein neues Konzept generiert hat (Programmzeile 20), ob dieses in dem vorangehenden, aber „falsch“ beantworteten Beispiel (:OLD.INSTANCE:) enthalten ist oder nicht (Programmzeile 30); wenn ja, wird nochmals ein neues Konzept generiert, wenn nein, wird das neue Konzept als das weiter zu verwendende Konzept beibehalten. Diese Modellvariante ist die einzige, die von der Konstruktion ‚make ‚OLD.INSTANCE“ :INSTANCE:‘ in SUBJECT’S.ANSWER (Abb. 5, Programmzeile 60) Gebrauch macht. Denkbar wäre jedoch auch eine Kombination dieses zusätzlichen Verarbeitungsschrittes mit dem GLOBAL.CONSISTENCY.MODEL (in dessen Programmzeile 30 einzubauen).

```
to LOCAL.NON.REPLACEMENT.MODEL
20 make „NEW.CONCEPT“
    GENERATE.CONCEPT from :MY.POSSIBLE.CONCEPTS:
30 if :NEW.CONCEPT = :MY.CONCEPT:
    then go to line 20
    else make „MY.CONCEPT“ :NEW.CONCEPT:
end
```

Abb. 9: Dritte Modellvariante der einfachen Begriffsbildung.

Das LOCAL.NON.REPLACEMENT.MODEL (Abb. 9) vergleicht nur noch, ob das neu generierte Konzept dem vorher verwendeten, aber falschen Konzept entspricht (Programmzeile 20); wenn ja, wird erneut ein Konzept generiert, wenn nein, wird es beibehalten.

```
to REPLACEMENT.MODEL
20 make „NEW.CONCEPT“
    GENERATE.CONCEPT from :MY.POSSIBLE.CONCEPTS:
30 make „MY.CONCEPT“ :NEW.CONCEPT:
end
```

Abb. 10: Vierte Modellvariante der einfachen Begriffsbildung.

Die simpelste Modellvariante ist das REPLACEMENT.MODEL (Abb. 10). In diesem wird lediglich ein neues Konzept generiert und dann beibehalten. Dabei kann es natürlich leicht vorkommen, daß das neu generierte Konzept

genau das gleiche wie das zuvor verwendete, aber bereits als falsch erwiesene Konzept ist, also eine wenig „intelligente“ Strategie einer Versuchsperson.

### 2.2.5 Abschließende Funktionsdefinitionen

Die in den bisher vorgestellten Unterprogrammen vorkommenden Teil- und Hilfsprogramme sind in den nachfolgenden Abb. 11-16 wiedergegeben.

```

to GENERATE.INSTANCE :CONCEPTS:
10 make „INSTANCE“ TAKE.ONE.OF.TWO from :CONCEPTS:
20 make „CONCEPTS“ butfirst of butfirst of :CONCEPTS:
30 if :CONCEPTS: = :empty:
    then output :INSTANCE:
40 make „INSTANCE“
    sentence of :INSTANCE:
    and TAKE.ONE.OF.TWO from :CONCEPTS:
50 go to line 20
end

```

Abb. 11: Teilprogramm zum Erzeugen eines Beispiels durch den Versuchsleiter.

```

to GENERATE.CONCEPT :CONCEPTS:
10 make „NUMBER“ random
20 if not :NUMBER: < count of :CONCEPTS:
    then go to line 10
30 if :NUMBER: = 0
    then output first of :CONCEPTS:
40 make „NUMBER“ :NUMBER: - 1
50 make „CONCEPTS“ butfirst of :CONCEPTS:
60 go to line 30
end

```

Abb. 12: Teilprogramm zum Erzeugen eines Konzepts durch die Versuchsperson.

```

to TAKE.ONE.OF.TWO :SENTENCE:
10 if (remainder of random and 2) = 0
    then output first of :SENTENCE:
    else output first of butfirst of :SENTENCE:
end

```

Abb. 13: Hilfsprogramm zur Ausgabe eines Elementes aus einem zweifach gegliederten Satz.

```

to CONTAINS :SENTENCE: :WORD:
10 if :SENTENCE: = :empty:
    then output „false“
20 if :WORD: = first of :SENTENCE:
    then output „true“
    else output CONTAINS butfirst of :SENTENCE: :WORD:
end

```

Abb. 14: Hilfsprogramm zum Prüfen auf Enthaltensein eines Wortes in einem Satz.

```

to REMOVE :WORD: :SENTENCE:
10 if :SENTENCE: = :empty:
    then output „ “
20 if :WORD: = first of :SENTENCE:
    then output butfirst of :SENTENCE:
    else output
        sentence of first of :SENTENCE:
        and REMOVE :WORD: from butfirst of :SENTENCE:
end

```

Abb. 15: Hilfsprogramm zum Entfernen eines Wortes aus einem Satz.

```

to from :ANYTHING:
10 output :ANYTHING:
end

```

Abb. 16: Definition des Füllwortes „from“.

Von psychologischer Bedeutung für die Simulation der einfachen Begriffsbildung sind nur noch die beiden Teilprogramme GENERATE.INSTANCE und GENERATE.CONCEPT. Mit GENERATE.INSTANCE (Abb. 11) wird ein Beispiel mit einer zufälligen Verteilung der beiden Werte des  $n$ -dimensionalen Reizes erzeugt, so daß jedes vom Versuchsleiter vorgelegte Beispiel eine Zufallsauswahl aus dem Reizmaterial darstellt (um beispielsweise Reihungseffekte zu vermeiden). Mit GENERATE.CONCEPT (Abb. 12) wird von der Versuchsperson eine Zufallsauswahl aus ihren möglichen Konzepten gemacht, wobei die psychologische Annahme zugrunde liegt, eine Bevorzugung bestimmter Reizdimensionen (wie z.B. Farbe oder Form) werde von der Versuchsperson nicht vorgenommen (eine empirisch durchaus widerlegbare Annahme).

Zum besseren Verständnis dieser Teilprogramme und der in Abb. 13-16 wiedergegebenen Hilfsprogramme (die ihrerseits nur noch von technischem

Interesse sind), seien noch einige Erläuterungen zur LOGO-Programmierung angefügt:

(1) Die LOGO-Operationen ‚first‘ und ‚butfirst‘ dienen zum *Zerlegen* von LOGO-Wörtern und LOGO-Sätzen; ‚first‘ liefert den ersten Buchstaben eines LOGO-Wortes bzw. das erste Wort eines LOGO-Satzes, ‚butfirst‘ liefert den Rest eines LOGO-Wortes ohne dessen ersten Buchstaben bzw. den Rest eines LOGO-Satzes ohne dessen erstes Wort.

(2) Die LOGO-Operation ‚random‘ liefert eine zufällige Zahl von 0-9, die LOGO-Operation ‚remainder‘ gibt den ganzzahligen Rest der Division zweier Zahlen aus, und die LOGO-Operation ‚count‘ zählt die Anzahl der Elemente eines LOGO-Wortes bzw. eines LOGO-Satzes.

(3) Im Unterschied zu allen anderen Programmen sind die beiden Hilfsprogramme CONTAINS (Abb. 14) und REMOVE (Abb. 15) *rekursiv* definierte Funktionen, d.h. sie rufen sich selbst innerhalb ihrer Definition wieder auf (in Programmzeile 20), bis die zugehörige Endbedingung erreicht ist (in Programmzeile 10).

(4) In Abb. 16 schließlich ist gezeigt, auf welch einfache Weise es möglich ist, in LOGO sog. Füllwörter („noise words“) zu schreiben, die zwar für den Programmablauf überflüssig sind (das Beispiel ‚from‘ ist nichts anderes als eine Identitätsoperation), für die Verständlichkeit einer Programmzeile jedoch von Nutzen sein können. (Anmerkung: Im Unterschied zu den bisherigen Konventionen ist ‚from‘ als benutzerdefinierte Operation nicht groß, sondern klein geschrieben, um das Schriftbild nicht mit solchen „noise words“ zu belasten.)

## 2.3 Diskussion des Programmbeispiels

### 2.3.1 Modellcharakteristika

Das im vorigen Abschnitt vorgestellte Programmbeispiel des „Simple Concept Attainment“ kann als Prototyp eines Simulationsmodells in der Psychologie angesehen werden:

- Es ist ein *dynamisches* Modell, denn es stellt den Zeitverlauf einer Experimentalsitzung in allen wesentlichen Aspekten dar - denen des Versuchsleiterverhaltens und insbesondere denen des Versuchspersonenverhaltens (des „Lernverhaltens“ der Versuchsperson).
- Es ist ein *deterministisches* Modell (mit probabilistischen Komponenten), da das Eingabe-Ausgabe-Verhalten des Modells eindeutig bestimmt und daher

vorhersagbar ist - bis auf die beiden Zufallsprozesse der Beispielgenerierung durch den Versuchsleiter (GENERATE.INSTANCE) und der Konzeptwahl durch die Versuchsperson (GENERATE.CONCEPT).

- Es ist ein qualitatives Modell insofern, als alle relevanten Modellvariablen eine nicht-numerische Spezifikation aufweisen, ohne jedoch eine weitergehende Quantifizierbarkeit des beobachtbaren Modellverhaltens auszuschließen.
- Es ist ein *analytisches* Modell, da von dem im realen Experiment beobachtbaren Gesamtverhalten ausgegangen wird, um das psychologisch zugrundeliegende Komponentenverhalten zu erschließen.
- Es ist ein *Erkundungsmodell* dahingehend, daß der Gegenstandsbereich der einfachen Begriffsbildung mit dem Hilfsmittel der Computer-Simulation in einer Weise untersucht werden kann, wie dies mit den herkömmlichen Mitteln der quantitativen Experimentalauswertung und der mathematischen Modellbildung nicht möglich ist. Gleichzeitig ist mit der Formulierung der verschiedenen Modellvarianten (GLOBAL.CONSISTENCY.MODEL, LOCAL.CONSISTENCY.MODEL, LOCAL.NON.REPLACEMENT.MODEL, REPLACEMENT.MODEL) auch die Möglichkeit gegeben, das Simulationsmodell im Sinne eines *Entscheidungsmodells* zu nutzen - das GLOBAL.CONSISTENCY.MODEL ist eine vergleichsweise optimale, das REPLACEMENT.MODEL die simpelste Variante der einfachen Begriffsbildung.
- Bezugspunkt des Simulationsmodells sind *interagierende Individuen*, nämlich Versuchsleiter und Versuchsperson. Will man sich allein auf das an der Versuchsperson beobachtbare Lernverhalten beschränken, da nur dieses einer psychologischen Erklärung bedarf, dann ist das isoliert betrachtete *Individuum* Bezugspunkt der Modellbildung.

### 2.3.2 Nicht-numerisches Programmieren

Hauptkennzeichen von Simulationsmodellen in der Psychologie ist, daß es sich um *nicht-numerische Modelle* handelt, für die eine ganz bestimmte Art des Programmierens charakteristisch ist, nämlich die des „nicht-numerischen Programmierens“. Nach Harbordt (1974, S. 41-42) ist ein nicht-numerisches Modell im einzelnen durch folgende Merkmale gekennzeichnet:

- (1) Der Gegenstandsbereich wird auf nicht-numerische Weise dargestellt, d.h. die Modellvariablen beschreiben den realen Prozeß oder das reale System in qualitativen Kategorien. (Das angeführte Programmbeispiel ist eine direkte Übersetzung der experimentellen Versuchsdurchführung in die nicht-numerische Programmiersprache LOGO.)

- (2) Die qualitativen Variablen werden in einer Hierarchie von „Listen“ angeordnet, und ihre Verarbeitung besteht in der Veränderung solcher Listen durch elementare Prozesse zum Sortieren, Ordnen, Speichern, Wiederaufsuchen, Vergleichen und Auswählen der jeweiligen Variablenwerte. (Diese Elementarprozesse werden in LOGO durch dessen Grundvokabular bereitgestellt.)
- (3) Der Modellablauf besteht in einer verschachtelten Abfolge von Programmen und Programmteilen (Haupt-, Unter-, Teil- und Hilfsprogramme des Programmbeispiels).
- (4) Die Modellausgabe sind in der Regel nicht-numerische, qualitative Daten, nämlich die Inhalte bestimmter Listen von Variablenwerten, die aufgrund der Programmdefinitionen zu eindeutig bestimmten Merkmalsklassen gehören. (Jede der Funktionsdefinitionen des Programmbeispiels beinhaltet eine genaue Spezifikation der jeweiligen Programmausgabe.)

Grundlage der Erstellung eines so charakterisierten Simulationsmodells ist jedoch eine andere Form des Programmierens, als man es durch die gängigen Programmiersprachen - wie z.B. ALGOL oder FORTRAN - gewohnt ist. Diese andere Form, das *nicht-numerische* Programmieren, ist nur mit speziell zu diesem Zweck entworfenen, sog. listenverarbeitenden Programmiersprachen möglich. Ein Beispiel dieser „nicht-numerischen“ Programmiersprachen ist die - hier als Illustrationssprache verwendete - Sprache LOGO (vgl. Feurzeig et al., 1971), eine andere die in den weitaus meisten Fällen der Erstellung von Simulationsmodellen und von Programmen der „künstlichen Intelligenz“ verwendete Programmiersprache LISP („LIST Programming language“, vgl. McCarthy et al., 1962).

Gemeinsames Merkmal aller nicht-numerischen oder listenverarbeitenden Programmiersprachen ist die Tatsache, daß der Computer in diesen Sprachen nicht eigentlich „rechnet“, sondern Daten beliebiger Struktur verarbeitet, sofern sie sich in digitalisierter Form (als aus zwei Grundwerten - z.B. 0/1 oder ON/OFF - bestehende „Bit-Folgen“) im Rechner darstellen lassen (wovon die Zahlen lediglich eine bestimmte Teilmenge bilden). Ihre konkrete Bedeutung erhalten diese Zeichen und Zeichenstrukturen erst durch die vom Benutzer eingeführten Funktionsdefinitionen, die über dem Grundvokabular einer Programmiersprache in deren „Grammatik“ realisierbar sind.

### 2.3.3 „Listenverarbeitung“

Grundlegendes Konzept des nicht-numerischen Programmierens ist das der *Listenverarbeitung*: Alle Zeichen und Zeichenstrukturen werden in Form von Listen dargestellt und verarbeitet, wobei es keinen prinzipiellen Unterschied

zwischen Programm und Daten gibt - Programm und Daten haben die gleiche Struktur, wenn auch im Programmablauf unterschiedliche Funktion. Das Konzept einer Listenstruktur läßt sich am klarsten am Beispiel der Programmiersprache LISP veranschaulichen. Eine *Liste* ist in LISP definiert als ein Klammerausdruck, dessen Elemente Wörter („Atome“ in LISP) sein können - oder aber weitere Listen („Unterlisten“). Mit dieser Definition eröffnet sich die Möglichkeit, Programme wie auch Daten als beliebig komplexe, verschachtelte Listen aufzubauen. Ein Beispiel ist das in Abb. 17 wiedergegebene, nunmehr in LISP formulierte Hauptprogramm des „Simple Concept Attainment“, das in Abb. 2 in der Programmiersprache LOGO vorgestellt wurde. (Zur besseren Verständlichkeit sind auch hier die vom Benutzer frei wählbaren Bezeichnungen groß und die zum LISP-Vokabular gehörigen Bezeichnungen klein geschrieben.)

```
(de SIMPLE-CONCEPT-ATTAINMENT
(TRUE-CONCEPT POSSIBLE-CONCEPTS CRITERION-VALUE
MODEL-VARIANT)
(prog (FEEDBACK RIGHT-ANSWER-COUNT)
      (setq FEEDBACK nil)
      (setq RIGHT-ANSWER-COUNT 0)
      LBL (print nil)
          (print (append (quote (INSTANCE IS A))
                        (EXPERIMENTER'S-INSTANCE)))
          (print (append (quote (**** I SAY)) (SUBJECT'S-ANSWER)))
          (print (append (quote (YOUR ANSWER IS))
                        (EXPERIMENTER'S-FEEDBACK)))
          (cond ((eq RIGHT-ANSWER-COUNT CRITERION-VALUE) (return))
                (t (go LBL))) ))
```

Abb. 17: Hauptprogramm zur einfachen Begriffsbildung in LISP.

Listenverarbeitung heißt nun, daß innerhalb einer Programmiersprache - hier LISP - jede Liste allein aufgrund ihrer Struktur und den in ihr vorkommenden Bezeichnungen eine bestimmte Bedeutung erhält. Am Beispiel von Abb. 17 ist die Listenstruktur anhand der Klammerung - und zur zusätzlichen optischen Verdeutlichung auch an den Einrückungen - klar erkennbar. Eine mit der LISP-Funktion ‚de‘ beginnende Liste erhält die Bedeutung einer Definitionsstruktur. Das zweite Element einer derartigen Struktur ist der vom Benutzer frei wählbare Name der zu definierenden Funktion. Als nächstes erwartet LISP eine Liste der Variablen oder Argumente, die später, beim Aufruf der Funktion, die einzugebenden Daten bezeichnen. Sodann folgt eine Liste, hier mit der LISP-Funktion ‚prog‘ eröffnet, die den eigentlichen Kern

der Definition enthält. Als erstes steht nach dem ‚prog‘ eine Liste von lokalen Variablen (die auch leer sein kann), die innerhalb des Definitionskerns lediglich eine lokale Bedeutung haben (im Unterschied zu den globalen, für alle möglichen Unterprogramme geltenden Variablen der Argumentliste nach dem Funktionsnamen). Die nachfolgende Sequenz von unterschiedlich tief geschachtelten Listen beschreibt den gleichen Programmablauf, hier in LISP-Terminologie, wie den in Abb. 2 in LOGO programmierten Versuchsablauf der einfachen Begriffsbildung. Wenn auch im Vergleich zwischen Abb. 2 und Abb. 17 der Begriff der Listenstruktur in dem LISP-Programm deutlicher zum Ausdruck kommt als in dem LOGO-Programm, so ist dieses doch - schon aufgrund seiner Wortwahl - wesentlich leichter zu verstehen. Zudem besteht auch in LOGO die Möglichkeit, eine der LISP-Notation ähnliche Klammer-schreibweise zu verwenden, um zusammengehörige Programmteile - insbesondere bei mehrfachen Funktionsaufrufen innerhalb einer Programmzeile - übersichtlicher zu gestalten; in manchen Fällen müssen in LOGO sogar Klammern geschrieben werden, wie beispielsweise in der LOGO-Operation TAKE.ONE.OF.TWO (Abb. 13) am Anfang von Programmzeile 10 (weitere Beispiele finden sich in den Abb. 20, 21, 25-28).

Auch die Datenstruktur, mit der das Programmbeispiel von Abb. 17 in LISP aufzurufen wäre, ist eine Listenstruktur, wie aus Abb. 18 zu ersehen ist. (Voraussetzung für einen Programmaufruf wäre natürlich noch eine den LOGO-Programmen von Abb. 4-16 entsprechende LISP-Programmierung.)

```
(SIMPLE-CONCEPT-ATTAINMENT
```

```
  (quote BLUE)
```

```
  (quote (SMALL BIG RED BLUE CIRCLE SQUARE))
```

```
5
```

```
  (quote GLOBAL-CONSISTENCY-MODEL))
```

Abb. 18: Aufrufbeispiel für das LISP-Programm der einfachen Begriffsbildung.

Aus dem rein syntaktischen Vergleich zwischen Abb. 17 und 18 ist ersichtlich, daß zwischen „Programm“ und „Daten“ kein grundsätzlicher Unterschied besteht. Der Unterschied ergibt sich erst durch den „Gebrauch“ der verschiedenen strukturierten und inhaltlich gefüllten Listen innerhalb der Programmiersprache LISP.

Nicht-numerisch ist das Programmieren in listenverarbeitenden Sprachen insofern, als numerische Werte - wie die ‚5‘ in Abb. 18 - nur einen Spezialfall unter den Daten eines so konzipierten Programms darstellen; in der Regel wird also nicht „gerechnet“, sondern es werden diskrete („digitalisierte“) Zeichen und Zeichenstrukturen in Listenform verarbeitet.



### 2.3.4 Modulares Programmieren

Aus dem Programmbeispiel des „Simple Concept Attainment“ wird ein wichtiges Konstruktionsprinzip von Simulationsmodellen erkennbar: das „modulare“ Programmieren (Modellkonstruktion nach dem „Baukastenprinzip“). Jedes der in den Abb. 2-16 wiedergegebenen LOGO-Programme („Haupt-, Unter-, Teil- und Hilfsprogramme“ in der dort eingeführten Terminologie) ist eine in sich selbständige Einheit - ein „Modul“ -, und erst aus dem verschachtelten Zusammenwirken dieser „Bausteine“ ergibt sich die Ablaufcharakteristik des Gesamtprogramms. Konkret heißt das für die Programmierarbeit an einem Simulationsmodell, daß man die einzelnen Programmeinheiten entweder „von oben nach unten“ („top-down“) oder „von unten nach oben“ („bottom-up“) erstellt.

Im „top-down programming“ schreibt man zuerst das Hauptprogramm (im Programmbeispiel: `SIMPLE.CONCEPT.ATTAINMENT`) und setzt zunächst nur die Namen von Unterprogrammen ein (Beispiele: `EXPERIMENTER'S.INSTANCE`, `SUBJECT'S.ANSWER`, `EXPERIMENTER'S.FEEDBACK`), über deren genauere Definitionsstruktur man sich hier noch keine Gedanken machen muß - es reicht eine vorläufige Vorstellung über die jeweiligen Ein- und Ausgaben dieser Unterprogramme. Bei der späteren Erstellung der Unterprogramme verfährt man in analoger Weise, bis am Ende alle für das Gesamtprogramm zu schreibenden Funktionen definiert sind. (Beispiel: Beim Schreiben des Unterprogramms `SUBJECT'S.ANSWER` setzt man die Namen der noch undefinierten Teilprogramme `GENERATE.CONCEPT` und `CONTAINS` so ein, daß man ein ganz bestimmtes Verhalten der Modellkomponente `SUBJECT'S.ANSWER` nach der Definition ihrer Teilprogramme erwarten kann. Im konkreten Fall dieses Beispiels ist insbesondere zu beachten, daß die Werte der globalen Variable `:MODEL.VARIANT`: - die ihrerseits im Hauptprogramm `SIMPLE.CONCEPT.ATTAINMENT` eingeführt wurde - später von LOGO nicht als Literale, sondern als auszuführende LOGO-Funktionen - durch die LOGO-Anweisung `'do'` veranlaßt - gelesen werden sollen.) Vorteil dieser Vorgehensweise ist die Möglichkeit, von zunächst relativ allgemeinen Vorstellungen eines Simulationsmodells zu immer konkreteren und spezifischeren Details überzugehen, ohne den Gesamtzusammenhang des Modellverhaltens aus dem Auge zu verlieren; in dieser Charakterisierung entspricht das „top-down programming“ der von Harbordt beschriebenen *analytischen* Modellentwicklung (vgl. Abschnitt 2.1, Punkt 4).

Im „bottom-up programming“ geht man den umgekehrten Weg: Von der Programmierung relativ spezifischer, meist unabhängig voneinander definierbarer Modellkomponenten gelangt man durch deren Einbau in allgemeinere Programme zu der Erstellung eines Gesamtprogramms, von dessen Verhalten man anfangs nur recht globale Vorstellungen haben muß. Erst durch die zunehmende Zusammenfassung von Modellkomponenten erfahren diese Vorstellun-

gen ihre konkrete Ausgestaltung, bis am Ende das erwünschte Gesamtverhalten des Simulationsmodells erreicht wird. Der Vorteil dieser Vorgehensweise liegt in der nahezu beliebig verfeinerbaren Herausbildung eines bestimmten Modellverhaltens; in der Terminologie von Harbordt entspricht diese Programmiertechnik der *synthetischen* Erstellung von Simulationsmodellen.

In der Programmierpraxis werden allerdings diese „reinen“ Formen der Modellerstellung eher die Ausnahme als die Regel sein; eine gemischte Strategie aus „top-down“ und „bottom-up“ Programmieren ist für das modulare Programmieren - insbesondere bei der Konstruktion komplexerer und umfangreicherer Simulationsmodelle - wohl mehr kennzeichnend. Seinen größten Vorteil hat das modulare Programmieren - im Unterschied zu einem alle Modellkomponenten in einem einzigen Programm enthaltenden Hauptprogramm - in der leichteren Fehlersuche („debugging“): Ist das Modellverhalten an einer bestimmten Stelle fehlerhaft - oder entspricht es nicht den intendierten Absichten -, so hat man nur die den Fehler bewirkende Modellkomponente (das entsprechende Unterprogramm) herauszufinden und zu verbessern (oder zu ersetzen), ohne die übergeordneten Programme oder gar das Hauptprogramm verändern zu müssen. Das erleichtert das Programmieren ganz wesentlich, eingedenk der alten Programmiererfahrung, daß jedes größere Programm so seine „Macken“ („bugs“) hat.

### 3. Simulationsmodelle und psychologische Theorienbildung

Im Unterschied zu dem empiristischen, oft nur mit ad-hoc-Hypothesen begründeten Vorgehen der quantitativ orientierten Methodik war die Entwicklung der Computer-Simulation als wissenschaftliche Methode von Anfang an mit einer expliziten psychologischen Theorienbildung verbunden, die die Verwendung des Rechners als Darstellungsmittel erst zu rechtfertigen gestattet. Die allgemeine Gestalt, die diese Theorienbildung angenommen hat, haben Newell & Simon (1972) in ihrem umfangreichen Buch über menschliches Problemlösen ausführlich beschrieben: Der Mensch als Gegenstand der Psychologie wird als ein *informationsverarbeitendes System* betrachtet, als ein mit seiner Umgebung in einem primär informationellen, nicht-materiellen Austausch befindliches System, vermittelt durch komplexe, intern gesteuerte Vorgänge des Wahrnehmens, Denkens, Fühlens und Verstehens, deren externe Beobachtbarkeit nur in eingeschränkter Weise gegeben ist - ein sehr ernst zu nehmen des empirisches Problem der kognitiven Psychologie. Nach Newell & Simon (1972, S. 9-13) ist die Theorie vom Menschen als einem informationsverarbeitenden System

(1) eine *Prozeßtheorie*, die von der Annahme einer begrenzten Anzahl intern wirksamer, das extern (oder an sich selbst) beobachtbare Verhalten produzierender Prozesse ausgeht;

- (2) eine Theorie des *Individuums*, in der individuelles Verhalten in spezifischen Einzelsituationen modelliert wird;
- (3) eine *inhaltsorientierte* Theorie, die die in ihren Modellen konkretisierten Aufgabenstellungen (wie z.B. Problemlösen) nicht nur zu beschreiben und zu erklären, sondern auch selbst auszuführen gestattet;
- (4) eine *dynamische* Theorie, in der ein Verhaltensablauf über der Zeit für jeden Handlungsschritt als eine Funktion des unmittelbar vorangehenden Zustands des Systems und seiner Umgebung dargestellt werden kann;
- (5) eine *empirische, nicht-experimentelle* Theorie dahingehend, daß einerseits so viele Daten wie nur möglich über die individuell verfügbare und tatsächlich verarbeitete Information benötigt werden, andererseits diese Daten jedoch nicht in dem reduzierten Bedingungsgefüge der herkömmlichen Experimentalpraxis zu gewinnen sind;
- (6) eine *nicht-statistische* Theorie, die - bis heute jedenfalls - von dem Apparat der Inferenzstatistik wenig Gebrauch machen kann, da die Daten und „Parameter“ der Modelle primär nicht-numerisch, qualitativ sind;
- (7) aber eine *hinreichende* Theorie in ihrer Fähigkeit, die kognitiven Phänomene, die sie untersucht, nicht nur am Menschen entdecken und beschreiben, sondern sie sogar in einem künstlichen System reproduzieren zu können.

Ob die psychologische Rahmentheorie, die mit der Computer-Simulation einhergeht, diese Merkmale aufweisen soll oder nicht, oder ob sie nicht eine ganz andere Charakterisierung erfahren sollte, kann und wird noch weiter diskutiert werden. Unabweisbar ist jedoch die Forderung, daß eine sinnvolle Verwendung der Computer-Simulation ohne eine begleitende psychologische Theorienbildung nicht wünschenswert ist.

## 3.1 Empirische Grundlagen psychologischer Simulationsmodelle

### 3.1.1 Methoden der Datengewinnung

Ausgangspunkt einer „inhaltsorientierten dynamischen Prozeßtheorie des Individuums“ - so die Charakterisierung in der Terminologie von Newell & Simon - ist eine empirische Datengewinnung, die die Eigenart und Vielfalt menschlicher Informationsverarbeitung widerzuspiegeln zumindest annäherungsweise gestattet. Wenig geeignet hierzu ist eine experimentalpsychologische Methodik, die vornehmlich an der Beobachtung isolierbarer Reaktionsweisen, an dem extern beobachtbaren Ergebnis von - möglicherweise sehr komplexen - internen Vorgängen orientiert ist, ohne sich ernsthaft die Frage zu stellen, ob nicht diese internen Prozesse ebenso untersuchenswert sind wie deren externe Resultate, auch wenn sie dem Experiment weniger zugänglich erscheinen. Nicht eine nur ergebnisorientierte, sondern eine mehr prozeßgeleitete empirische Datengewinnung kann den Zugang zu den psychologischen Phänomenen der menschlichen Informationsverarbeitung eröffnen.

Als Methode der Wahl hat die kognitive Psychologie auf die schon von Wundt eingeführte, heute jedoch in liberalisierter Form gehandhabte „Introspektion“ zurückgegriffen, jene Form selbstexplorativen Verhaltens, die dem Behaviorismus stets wissenschaftlich verdächtig gewesen ist. Diese liberalisierte Variante der Introspektion ist die sog. *Methode der „lauten Denkens“*, des spontanen, sich frei entwickelnden Verbalisierens von Inhalten und Vorgängen des Bewußtseins beim Denken, Lernen, Wahrnehmen, Verstehen, Handeln usw. Typischerweise sieht eine derartige Datengewinnung so aus, daß einem Probanden eine Aufgabenstellung, sei es in einer Versuchssituation oder in einer alltagsnahen Umgebung, vorgegeben ist, deren Lösung er nicht still in seinem Kopf, sondern durch ein begleitendes „lautes Denken“ Schritt für Schritt entwickelt, wobei sein externalisiertes Verhalten einschließlich seines Verbalisierungsverhaltens auf Videoband oder Tonband aufgezeichnet wird. Ergebnis ist ein „Protokoll“, das - ganz im Sinne von Punkt (4) der Charakterisierung von Newell & Simon - den „Verhaltensablauf über der Zeit für jeden Handlungsschritt als eine Funktion des unmittelbar vorangehenden Zustands des Systems und seiner Umgebung“ beinhaltet, wenn auch noch ganz im Sinne von nicht-analysierten „Rohdaten“.

Als Beispiel für ein derartiges „Rohdatenprotokoll“ ist in Tabelle 1 das Verbalisierungsprotokoll eines Probanden wiedergegeben, der Intelligenztestaufgaben vom Typ des „Unpassenden Streichens“ mit der Methode des „lauten Denkens“ zu lösen hatte: Gegeben seien die Buchstabengruppen

AABC ACAD ACSH AACG;

welches ist die unpassende Buchstabengruppe, die nicht zu den anderen paßt?

Tabelle 1: Verbalisierungsprotokoll eines Probanden beim Lösen einer Aufgabe des „Unpassenden Streichens“.

- 1 „Also, wir haben vier Buchstabengruppen hier:
- 2 AABC, ACAD, ACSH und AACG.
- 3 Woll'n zuerst mal seh'n, welche von denen gleiche Buchstaben drin haben:
- 4 AABC hat zwei A's,
- 5 ACAD auch,
- 6 ACSH nicht;
- 7 aha!
- 8 Aber AACG wieder.
- 9 Also ist ACSH die unpassende Gruppe, paßt nicht zu den andern.“

Das Verbalisierungsprotokoll in Tabelle 1 wurde der Übersichtlichkeit halber schon in zusammengehörige Segmente gegliedert und zur leichteren Orientierung mit einer Zeilennumerierung versehen. Im Grunde genommen ist diese

Segmentierung - und sogar die Übertragung des Tonbandprotokolls in die schriftliche Form mit entsprechender Interpunktion - schon ein erster Schritt der Datenauswertung (auf deren Aspekte im nächsten Abschnitt eingegangen werden wird). Die 9 Protokollzeilen geben zwar sicher kein vollständiges Bild der in dem Probanden insgesamt ablaufenden Prozesse des Problemlösens, sind aber doch informativ genug, um sich eine Vorstellung von dem Lösungsprozeß derartiger Aufgaben zu machen - eine bessere Vorstellung jedenfalls, als wenn man von dem Probanden nur die Angabe „ACSH“ in Protokollzeile 9 bekäme, wie dies bei der üblichen Durchführung von Intelligenztests der Fall ist.

Neben der speziellen Verbalisierungsmethode des „lauten Denkens“ sind natürlich auch alle anderen Methoden der Datengewinnung verwendbar, die eine Aufzeichnung von Sprachverhalten - als unmittelbarer Ausdruck von Vorgängen der Informationsverarbeitung - ermöglichen. Dazu zählen beispielsweise Befragungstechniken wie das freie oder das standardisierte Interview, Gruppendiskussionen in Problemlöse- und Entscheidungssituationen, Interaktions- und Gesprächsverläufe von therapeutischen Sitzungen, ja selbst das Geschehen in Encounter-Gruppen könnte die empirische Grundlage für ein Simulationsmodell von „selbstexplorativen Gruppenvorgängen“ - so das hier zugrunde zu legende theoretische Konzept - liefern.

Darüber hinaus können durchaus auch Datenquellen der herkömmlichen Experimentalforschung herangezogen werden, sofern sie nicht nur ergebnisorientiert sind, sondern auch die Protokollierbarkeit von Prozeßabläufen beinhalten. Ein Beispiel ist die *Blickbewegungsregistrierung* einer Versuchsperson in Experimenten, die ein großflächig projizierbares Bildmaterial - wie z.B. Matrizenaufgaben aus Intelligenztests oder Schachpositionen - verwenden. Die Aufzeichnung der Blickbewegungen erfolgt mit einer speziell dazu konstruierten „Eye Marker“-Kamera, meist noch verbunden mit einer Tonbandaufzeichnung des Verbalisierungsverhaltens der Versuchsperson (vgl. Newell & Simon, 1972).

### 3.1.2 Möglichkeiten der Datenauswertung

In den meisten Fällen dient die empirische Datenerhebung der vorbereitenden Phase der Erstellung eines Simulationsmodells; aber auch in der abschließenden Phase der Modellprüfung ist auf die empirische Basis zurückzugreifen, um über die Validität des Modells etwas aussagen zu können (vgl. Abschnitt 4, Validierung und Anwendbarkeit von Simulationsmodellen). Die Phase der Modellerstellung auf der Grundlage empirischer Daten ist ein komplizierter - und bei den Datenmengen von Verbalisierungsprotokollen ein aufwendiger - Vorgang der Datenauswertung, der ohne ein theoriegeleitetes Arbeiten am

Material kaum durchführbar ist. Eine rein empiristische, allenfalls durch ad-hoc-Hypothesen angereicherte Datenauswertung ist bei der Vielfalt und dem Reichtum von Prozeßdaten schnell zum Scheitern verurteilt, da selbst das umfangreichste Datenmaterial hinsichtlich der Vorgänge, deren Abbild es ist, so lückenhaft sein kann, daß ohne eine theoriegestützte Ergänzung ein vollständiges Bild des Prozeßgeschehens nicht erreichbar ist.

Die Auswertung von Verbalisierungsprotokollen, für die es auch schon computergestützte Verfahren gibt (vgl. Simon, 1979), steht im Vordergrund der psychologischen Datenanalyse bei der Erstellung eines Simulationsmodells. Ziel der *sog. Protokollanalyse* ist die detaillierte Aufschlüsselung des den Verbalisierungsdaten zugrundeliegenden Prozesses der Informationsverarbeitung, zu dessen Rekonstruktion das Simulationsmodell erstellt werden soll. Die *psychologische Rahmentheorie*, in deren Kontext die Protokollanalyse ihre Grundlage hat, läßt sich - in Verallgemeinerung der am Beispiel des Problemlösens entwickelten Theorie von Newell & Simon (1972, Kap. 14) - auf zwei Annahmen über die allgemeine Natur der menschlichen Informationsverarbeitung aufbauen:

(1) Zu jedem Zeitpunkt befindet sich das informationsverarbeitende System (Mensch, Computer) in einem bestimmten *Kenntniszustand*, dessen interne Darstellung die Form von Zeichen und Zeichenstrukturen („Symbols“, „symbol structures“) hat, die externe Gegebenheiten in der Umgebung des Systems (Dinge, Ereignisse, Vorgänge) oder interne, das System selbst betreffende Sachverhalte (Wahrnehmungen, Gedanken, Erinnerungen, Empfindungen, Stimmungen - wenn auch in dieser Terminologie nicht so sehr auf einen Computer zutreffend!) abbilden.

(2) Jeder Kenntniszustand wird durch die geeignete Anwendung eines bestimmten *kognitiven Operators* in einen anderen Kenntniszustand überführt, so daß das gesamte Geschehen in einem informationsverarbeitenden System aus der Abfolge der einzelnen Operationen vollständig beschrieben werden kann; jeder kognitive Operator ist durch die Angabe seiner als Eingabe dienenden und seiner als Ausgabe resultierenden Kenntniszustände definierbar. Die *Elementaroperationen* kognitiver Operatoren sind im einzelnen (vgl. auch Newell & Simon, 1972, S. 29-30; von den Autoren als „elementare Informationsprozesse“ bezeichnet):

- (a) Aufnehmen von Information aus der Systemumgebung („Sinneswahrnehmung“) bzw. aus dem Systeminnern („Befindlichkeiten“) und deren interne Repräsentation als Zeichen und/oder Zeichenstrukturen;
- (b) Abgeben von Information an geeignete Effektororgane des Systems (Sprechen, Schreiben, manuelle Tätigkeiten, Körperfunktionen im Bereich der menschlichen Informationsverarbeitung);
- (c) Speichern von Information in verschiedenen Speichermedien (Kurz- und

Langzeitgedächtnis, ggf. auch „sensorische Speicher“ der Sinnessysteme des Menschen);

(d) Erkennen von Information als im Arbeitsspeicher („Kurzzeitgedächtnis“) repräsentierte Zeichen und Zeichenstrukturen;

(e) Vergleichen von Information hinsichtlich Gleichheit/Ähnlichkeit/Verschiedenheit von Zeichen und Zeichenstrukturen;

(f) Erzeugen von Information durch Zusammensetzen von Zeichenstrukturen aus einzelnen Zeichen oder Teilstrukturen bzw. durch Zerlegen von Zeichenstrukturen in ihre Bestandteile (Zeichen, Teilstrukturen);

(g) Löschen von für die weitere Verarbeitung nicht mehr benötigter Information in den verschiedenen Speichermedien.

Wie diese rahmentheoretischen Vorstellungen über die Grundlagen eines informationsverarbeitenden Systems die Protokollanalyse von Verbalisierungsdaten anzuleiten gestatten, sei am Beispiel des in Tabelle 1 wiedergegebenen Probandenprotokolls illustriert. Eine Analyse dieses Protokolls hat für jede verbalisierte Äußerung zu zeigen, von welchem Kenntniszustand ausgehend ein nachfolgender Kenntniszustand durch die Anwendung ganz bestimmter kognitiver Operatoren erzeugt worden sein kann. Das Ergebnis einer derartigen Analyse ist zunächst noch rein hypothetisch und erst die nachfolgende Erstellung eines entsprechenden Simulationsmodells kann die Angemessenheit der Datenauswertung sichtbar machen. Bei genauerer Durchsicht des Verbalisierungsprotokolls von Tabelle 1 kann man annehmen, daß in den einzelnen Kenntniszuständen des Probanden Konzepte wie „Buchstabengruppe“, „Unpassende Gruppe“ und das Lösungskonzept „Gleiche Buchstaben“ eine Rolle gespielt haben. Als Operatoren mögen dem Probanden kognitive Operationen wie „Lesen“ (von Buchstabengruppen), „Suchen“ (nach einem Lösungskonzept), „Verwenden“ (des gefundenen Lösungskonzeptes), „Merkens“ (der unpassenden Gruppe) und „Beantworten“ (der Aufgabenstellung) zur konkreten Gestaltung seines Problemlöseprozesses zur Verfügung gestanden haben. Operationen also, die sich teils als Elementaroperationen und teils aus solchen zusammengesetzt interpretieren lassen. Die Rekonstruktion des Gesamtprozesses der Informationsverarbeitung am Beispiel dieser Intelligenztestaufgabe ist in Tabelle 2 wiedergegeben, aus der nicht nur die genaue Bedeutung der oben angeführten Konzepte und Operationen ersichtlich wird, sondern auch deren jeweilige Zuordnung zu den entsprechenden Ausschnitten aus dem Verbalisierungsprotokoll des Probanden. Bemerkenswert ist hierbei, daß selbst ein so klares und schlüssiges wie das hier mitgeteilte Protokoll „Verbalisierungslücken“ aufweist (vgl. die Leerstellen im Protokollteil von Tabelle 2), die sowohl „entdeckt“ als auch „geschlossen“ werden können nur durch die begleitenden rahmentheoretischen Vorstellungen über die Stringenz eines zielführenden Prozesses der Informationsverarbeitung. Deren Begründbarkeit kann dann allerdings erst die nachfolgende Erstellung eines Simulationsmodells liefern.

Die Notation in Tabelle 2 wurde bereits so gewählt, daß eine eventuelle Programmierung in LOGO erleichtert wird: Die Operatoren wären als Funktionen zu definieren und die in die Kenntniszustände eingehenden Konzepte als Variablen, deren Werte Literale sind (in diesem Fall LOGO-Wörter und -Sätze). Zu beachten ist, daß alle Ein- und Ausgaben der Operatoren sich auf Inhalte des sog. Arbeitsspeichers („Kurzzeitgedächtnis“) beziehen, so daß in manchen Fällen die Eingabe (bei dem LESE-Operator) und gegebenenfalls auch die Ausgabe leer sein kann. Die „Programmlogik“ der Informationsverarbeitung an diesem Beispiel einer Intelligenztestaufgabe ist aus der Darstellung von Tabelle 2 klar erkennbar: Nach dem Einlesen der Buchstabengruppen wird zunächst nach einem Lösungskonzept („GLEICHE BUCHSTABEN“) gesucht. Danach wird dieses (mittels des Operators VERWENDE :LÖSUNGSKONZEPT:) auf die einzelnen Buchstabengruppen angewendet, wobei jede Buchstabengruppe erneut eingelesen wird (vgl. LESE :BUCHSTABENGRUPPE:, wofür es in dem Verbalisierungsprotokoll keine Hinweise gibt; hier könnte eine Blickbewegungsregistrierung die Verbalisierungslücken überbrücken helfen). Wie ersichtlich wird, ist nur auf die Buchstabengruppe „ACSH“ das Lösungskonzept nicht zutreffend; also ist diese die :UNPASSENDE GRUPPE:, die am Ende als Beantwortung der Aufgabenstellung auch ausgegeben wird.

Die tatsächliche Erstellung eines Simulationsmodells für Aufgaben des „Unpassenden Streichens“ müßte die Stringenz des oben gezeigten Programmablaufs deutlich zum Ausdruck bringen; insbesondere sollte das Modell auch eine hinreichende Begründung für das Ausfüllen von Verbalisierungslücken liefern, wie dies für das obige Beispiel vorgenommen wurde.

Andere Methoden der Analyse von Verbaldaten - wie z.B. die in Soziologie und Politologie verwendete Methode der „Inhaltsanalyse“ zur Auswertung von Textmaterial hinsichtlich quantitativ-statistischer Zusammenhänge (Themen- und Worthäufigkeiten und Korrelationen darüber) - spielen für die Computer-Simulation in der Psychologie nur eine untergeordnete Rolle.

Von zunehmender Bedeutung ist dagegen die Kombination von Auswertungsmethoden wie beispielsweise die mit der Protokollanalyse von Verbalisierungsdaten verknüpfbare Auswertung von Blickbewegungsdaten. Nicht nur, daß hierbei die Wahrnehmungskomponente stärker in die Modellierung der menschlichen Informationsverarbeitung einbezogen werden kann, ist der Vorteil dieser Methodenkombination, sondern auch, daß damit gezeigt werden kann, inwieweit die „verbalisierte“ Information mit der „visualisierten“ kongruent geht, oder ob - was zu vermuten wäre - letztere nicht vielmehr ersteren vorausseilt. Das in Tabelle 2 wiedergegebene Beispiel einer Protokollanalyse wäre - wie schon angedeutet - empirisch mit einer Analyse von Blickbewegungsdaten sicher leichter abzusichern. - Eine noch „ganzheitlichere“ Darstellung von Prozessen der Informationsverarbeitung ließe sich



Tabelle 2: Protokollanalyse der Verbalisierungsdaten eines Probanden beim Lösen einer Aufgabe des „Unpassenden Streichens“.

Operatoren (OP) und deren Eingaben (E) und Ausgaben (A)	Protokollausschnitt
OP (LESE :BUCHSTABENGRUPPEN:) E A (:BUCHSTABENGRUPPEN: „AABC ACAD ACSH AACG“)	„Also, wir haben vier Buchstaben-gruppen hier: AABC, ACAD, ACSH und AACG.“
OP (SUCHE :LÖSUNGSKONZEPT:) E (:BUCHSTABENGRUPPEN: „AABC ACAD ACSH AACG“) A (:LÖSUNGSKONZEPT: „GLEICHE BUCHSTABEN“)	„Woll’n zuerst mal seh’n, welche von denen gleiche Buch-staben drin haben:“
OP (LESE :BUCHSTABENGRUPPE:) E A (:BUCHSTABENGRUPPE: „AABC“)	
OP (VERWENDE :LÖSUNGSKONZEPT:) E (:BUCHSTABENGRUPPE: „AABC“) A (:GLEICHE BUCHSTABEN: „JA, ZWEI A“)	„AABC hat zwei A’s,”
OP (LESE :BUCHSTABENGRUPPE:) E A (:BUCHSTABENGRUPPE: „ACAD“)	
OP (VERWENDE :LÖSUNGSKONZEPT:) E (:BUCHSTABENGRUPPE: „ACAD“) A (:GLEICHE BUCHSTABEN: „JA, ZWEI A“)	„ACAD auch,“
OP (LESE :BUCHSTABENGRUPPE:) E A (:BUCHSTABENGRUPPE: „ACSH“)	
OP (VERWENDE :LÖSUNGSKONZEPT:) E (:BUCHSTABENGRUPPE: „ACSH“) A (:GLEICHE BUCHSTABEN: „NEIN, KEINE“)	„ACSH nicht;“
OP (MERKE :BUCHSTABENGRUPPE:) E (:BUCHSTABENGRUPPE: „ACSH“) (:GLEICHE BUCHSTABEN: „NEIN, KEINE“) A (:UNPASSENDE GRUPPE: „ACSH“)	„aha!“

Tabelle 2: Fortsetzung

Operatoren (OP) und deren Eingaben (E) und Ausgaben (A)	Protokollausschnitt
OP (LESE :BUCHSTABENGRUPPE:) E A (:BUCHSTABENGRUPPE: „AACG“)	
OP (VERWENDE :LÖSUNGSKONZEPT:) E (:BUCHSTABENGRUPPE: „AACG“) A (:GLEICHE BUCHSTABEN: „JA, ZWEI A“)	„Aber AACG wieder.“
OP (BEANTWORTE :AUFGABENSTELLUNG:) E (:UNPASSENDE GRUPPE: „ACSH“) A (:UNPASSENDE GRUPPE: „ACSH“)	„Also ist ACSH die unpassende Gruppe, paßt nicht zu den ändern.“

erzielen, könnte man das gesamte, beispielsweise auf Videoband aufgezeichnete Verhalten einer Person in die Datenauswertung einbeziehen: Verbalverhalten wäre sinnvoll durch nichtverbales Verhalten ergänzt. Allerdings - die Kategorien und Verfahren für eine derartige „ganzheitliche“ Datenauswertung sind erst noch zu entwickeln.

### 3.2 Informationelle Produktionssysteme

Wesentliches Merkmal der in Verbindung mit der Computer-Simulation entwickelten psychologischen Theorienbildung ist die Darstellungsbreite, mit der der jeweilige Gegenstandsbereich abgebildet wird. Vor allem auffallend ist der Versuch, die Prozeßtheorie der menschlichen Informationsverarbeitung in ihren unterschiedlichen Anwendungsbereichen mit Strukturtheorien des Gedächtnisses zu verknüpfen (vgl. Wender, Colonius & Schulze, 1980), wie es beispielsweise die neueren Theorien zum Sprachverstehen von Anderson & Bower (1973), Norman, Rumelhart & LNR (1975), Anderson (1976) und Schank & Abelson (1977), aber auch die „Vorläufer-Theorie“ zum menschlichen Problemlösen von Newell & Simon (1972) zeigen.

Die in ihrer Bedeutung wohl umfassendste Systemarchitektur des menschlichen Gedächtnisses und seiner Informationsverarbeitung baut auf der von Newell & Simon (1972) entwickelten Konzeption „*informationeller* (oder kognitiver) *Produktionssysteme*“ auf (vgl. Hunt & Poltrock, 1974: Ueckert, 1980a). Von ihrer Verwendungsweise her betrachtet sind informationelle Produktionssysteme die „Assembler-Sprache“ der menschlichen Informationsver-

arbeitung, in der - sollte die Theorie sich als empirisch zutreffend durchsetzen - der „informationelle Kode“ kognitiver Aktivität geschrieben ist, der dann in Form von konkreten Simulationsmodellen auf dem Rechner nachgebildet werden kann. Mit diesem Anspruch hat die Produktionssystem-Konzeption nicht nur in die Computer-Simulation kognitiver Prozesse Eingang gefunden, sondern auch in die „künstliche Intelligenz“-Forschung (vgl. beispielsweise den Sammelband von Waterman & Hayes-Roth, 1978).

### 3.2.1 Die Modellarchitektur von Produktionssystemen

Grundlegendes Konzept der Modellarchitektur von Produktionssystemen ist der Begriff der *Produktionsregel* (oder kurz: Produktion). Eine Produktionsregel beschreibt den Sachverhalt, unter welchen *Konditionen*  $K$  (d.h. bei welchem gegebenem Kenntniszustand des informationsverarbeitenden Systems) welche *Aktionen*  $A$  (d.h. welche kognitiven Operationen) ausgeführt werden sollen, um einen neuen Kenntniszustand des Systems zu erreichen.

Ein einfaches Beispiel sind die beiden Produktionsregeln:

$FA_1$  „Ampel ist rot“  $\rightarrow$  Warten, Ampel beobachten

$FA_2$  „Ampel ist grün“  $\rightarrow$  Gehen

Formal hat eine Produktionsregel stets die Struktur

$$N \ K \rightarrow A,$$

wobei  $K$  den Konditionalteil,  $A$  den Aktionsteil und der „Übergangspfeil“  $\rightarrow$  die Kopplung von  $A$  an  $K$  bezeichnet;  $N$  ist der „Name“ der Produktionsregel.

Ein *Produktionssystem* ist dann eine Menge (Liste) von Produktionsregeln, die in bestimmter Weise abgearbeitet werden. Obiges Beispiel der beiden Produktionen  $FA_1$  und  $FA_2$  kann man als Produktionssystem für das Verhalten an einem Fußgängerüberweg mit Ampelregelung ansehen, in dem aufgrund bestimmter Gegebenheiten („Ampel ist rot“ oder „grün“) bestimmte Handlungen (Warten und Ampel beobachten oder Gehen) ausgeführt werden (wenn auch in diesem Falle keine „kognitiven Operationen“ vorliegen, sondern motorische Aktivitäten, deren „Programmierung“ man sich jedoch als entsprechende Produktionssysteme vorstellen kann).

Zur Modellarchitektur von informationellen Produktionssystemen gehören - sowohl in ihrer Realisierung im Menschen als auch für deren Simulation auf einem Rechner - die folgenden drei *Systemkomponenten*:

(1) Ein oder mehrere *Arbeitsspeicher*, in denen alle augenblicklich verfügbare Information - aus welchen Quellen auch immer - kurzzeitig gespeichert

wird; psychologisch betrachtet sind der oder die Arbeitsspeicher das menschliche Kurzzeitgedächtnis und die unterschiedlichen „sensorischen Speicher“.

(2) Ein *Produktionsspeicher*, in dem die zu Produktionssystemen zusammengefaßten Produktionsregeln langfristig verfügbar sind und nach den jeweiligen Anforderungen der Informationsverarbeitung aktiviert, aber auch modifiziert und gelöscht werden können; die psychologische Instanz für den Produktionsspeicher ist das menschliche Langzeitgedächtnis.

(3) Ein *Interpreter*, der in der Lage ist, sowohl die Inhalte der Arbeitsspeicher (d.h. deren „Daten“) als auch die des Produktionsspeichers (d.h. dessen „Regeln“) zu „lesen“ und entsprechend dem jeweils aktivierten Produktionssystem zu „handeln“. Psychologisch gesehen ist der Interpreter von Produktionssystemen der „zentrale Prozessor“ oder die „kognitive Exekutive“ und kann somit durchaus als Konstrukt für das menschliche Bewußtsein verstanden werden (vgl. Ueckert, 1980b): Hauptmerkmal unseres Bewußtseins ist die Fähigkeit, die Aufmerksamkeitsverteilung über die Bewußtseinsinhalte so zu regeln, daß ein zielgerichtetes Verhalten des Gesamtsystems sowohl intern (in der Informationsverarbeitung selbst) als auch extern (in dem von außen beobachtbaren individuellen Handeln) resultiert.

### 3.2.2 *Beispiel eines Produktionssystems als Simulationsmodell*

Als Einführung in Konzeption und Arbeitsweise von informationellen Produktionssystemen sei im folgenden ein Produktionssystem vorgestellt, das als ein Simulationsmodell für die Aufgabe des „Unpassenden Streichens“ („Single Letter Exclusion“) angesehen werden kann, wobei auf die Verbalisierungsdaten zu dieser Aufgabe (vgl. Tabelle 1) sowie auf deren Protokollanalyse (vgl. Tabelle 2) für eine Diskussion des Modells zurückgegriffen werden kann. Wünschenswert ist eine Modellentwicklung, in der der konkrete Ablauf der Informationsverarbeitung so detailliert verfolgt werden kann, daß ein direkter Vergleich mit den Verbalisierungsdaten und der Protokollanalyse möglich ist.

Das Modellbeispiel ist - wie die bisherigen Programmbeispiele - in der Notation der Programmiersprache LOGO formuliert, so daß eine Übertragung in ein lauffähiges Computer-Programm unmittelbar gegeben ist. Zur Realisierung der Modellarchitektur wird als Arbeitsspeicher auf den Variablenspeicher („Namenspeicher“) von LOGO zurückgegriffen, während als Produktionsspeicher das von LOGO mit der ‚get‘-Anweisung aktivierbare Datei-System (langfristiges Speichersystem) verwendet wird. Der Interpreter ist als ein sequentielles LOGO-Programm konzipiert. Die Darstellung der Produktionsregeln wird einheitlich in der Form

$$\begin{array}{l} N \\ K \\ \rightarrow A \end{array}$$

durchgeführt (wobei N den Namen, K den Konditionalteil und A den Aktionsteil der Produktionsregel bezeichnet).

In Abb. 20 sind die für das Produktionssystem „Single Letter Exclusion“ benötigten Produktionsregeln wiedergegeben. Bevor auf sie inhaltlich eingegangen werden kann, sollen die neu vorkommenden LOGO-Ausdrücke kurz erläutert werden: Die LOGO-Operation ‚request‘ ist eine Anfrage an den Benutzer (am Terminal durch das Ausdrucken eines Sterns \* angezeigt), dem Programm eine Eingabe einzutippen. Der Ausdruck ‚thing‘ ist eine LOGO-Operation, die den Wert („das Ding“) einer Variable liefert (normalerweise wird diese Operation nicht benötigt, da man durch Aufruf der Variable deren Wert bekommt; ist jedoch dieser Wert selbst wieder eine Variable, so kann man zu deren Wert mit der Operation ‚thing‘ zugreifen). Der Ausdruck ‚both‘ ist die logische Und-Funktion (Konjunktion), d.h. ‚both‘ liefert den Wert „true“, wenn die beiden in der Konjunktion stehenden Prädikate den Wert „true“ haben. Die in LOGO vorgegebene Variable :bell: hat das am Terminal vorhandene Klingelsignal als ihren Wert.

Die Programmierung der Produktionsregeln erfolgt in LOGO in Form von Funktionsdefinitionen: Man schreibt vor den Namen einer Produktionsregel die LOGO-Anweisung ‚to‘, beginnt den Konditionalteil mit dem Ausdruck ‚10 test‘ (‚test‘ ist eine alternative LOGO-Operation zu der ‚if-then-else‘-Konstruktion), ersetzt den Übergangspfeil  $\rightarrow$  durch den Ausdruck ‚20 iftrue‘ (‚iftrue‘ entspricht dem ‚then‘ in ‚if-then-else‘) und beendet die Funktionsdefinition mit der LOGO-Anweisung ‚end‘.

Die konkrete Erstellung von Produktionsregeln folgt ganz dem Prinzip des „bottom-up programming“ und ist damit ein Beispiel für das modulare Programmieren bei der Entwicklung eines Simulationsmodells (vgl. Abschnitt 2.3.4): Jede Produktionsregel ist eine selbständige Einheit, die von allen anderen Produktionsregeln unabhängig ist (d.h. Produktionen können sich wechselseitig nicht aufrufen). Das „bottom-up programming“ wird wesentlich erleichtert, wenn empirische Daten (wie im vorliegenden Fall beispielsweise ein Verbalisierungsprotokoll und dessen Analyse) gegeben sind, die die Formulierung einzelner Produktionsregeln anzuleiten gestatten. Die in Abb. 20 wiedergegebenen Produktionsregeln zum „Single Letter Exclusion“ sind so auch leichter zu verstehen, wenn sie in direktem Vergleich zu der in Tabelle 2 dargestellten Protokollanalyse gelesen werden. Die ersten vier Produktionen entsprechen ziemlich genau den ersten vier Abschnitten der Protokollanalyse; sie beschreiben das Einlesen von Buchstabengruppen (SILEX1 bzw. SILEX3), die Suche nach einem Lösungskonzept (SILEX2) und dessen Anwendung auf

SILEX1

```
:GIVEN.ITEMS: = :empty:
→ make „GIVEN.ITEMS“ request ,
  get LETTER CONCEPTS
```

SILEX2

```
:CONCEPT: = :empty:
→ make „CONCEPT“ first of :LETTER.CONCEPTS:
```

SILEX3

```
:ITEM: = :empty:
→ make „ITEM“ request
```

SILEX4

```
ACTIVE :ITEM:
→ do :CONCEPT:
```

SILEX5

```
(first of thing of :CONCEPT:) = „YES“
→ make thing of „CONCEPT“ :empty: ,
  make „ITEM“ :empty:
```

SILEX6

```
(first of thing of :CONCEPT:) = „NO“
→ make „UNSUITABLE.ITEM“
  sentence of :ITEM: and :UNSUITABLE.ITEM: ,
  make thing of „CONCEPT“ :empty: ,
  make „ITEM“ :empty:
```

SILEX7

```
both (count of :UNSUITABLE.ITEM:) = 1 and :ITEM: = :bell:
→ print :UNSUITABLE.ITEM: ,
  make „HALT“ „PROBLEM IS SOLVED“
```

SILEX8

```
both not (count of :UNSUITABLE.ITEM:) = 1 and :ITEM: = :bell:
→ make „UNSUITABLE.ITEM“ :empty: ,
  make „LETTER.CONCEPTS“
  REMOVE :CONCEPT: from :LETTER.CONCEPTS: ,
  make „ITEM“ :empty:
```

Abb. 20: Produktionsregeln für die Aufgabe des „Unpassenden Streichens“ („Single Letter Exclusion“).

die jeweils zu bearbeitende Buchstabengruppe (SILEX4). Für die übrigen Produktionen läßt sich in der Protokollanalyse nicht immer ein direktes Analogon finden, doch ist ihre psychologische Plausibilität einsichtig: Produktionen SILEX5 und SILEX6 beschreiben das Problemlöseverhalten nach Anwendung

des Lösungskonzeptes und Produktionen SILEX7 und SILEX8 das entsprechende Verhalten nach Bearbeitung aller vorgegebenen Buchstabengruppen (durch das Klingelsignal :bell: angezeigt); mit SILEX7 wird das Problemlösen erfolgreich abgeschlossen, mit SILEX8 jedoch fortgesetzt, nachdem sich das gewählte Lösungskonzept als unbrauchbar erwiesen hat (d.h. der ganze Prozeß wiederholt sich mit der Suche nach einem neuen Lösungskonzept).

Zur Arbeitsweise einzelner Produktionsregeln sind einige Erläuterungen angebracht:

(1) Die LOGO-Anweisung ‚get LETTER CONCEPTS‘ im Aktionsteil von SILEX1 setzt das Vorhandensein einer Datei unter dem Namen LETTER CONCEPTS voraus, in der die für Buchstabenaufgaben verwendbaren Lösungskonzepte und deren Funktionsdefinitionen langfristig gespeichert sind. Im konkreten Beispiel der vorliegenden Aufgabe habe die Datei LETTER CONCEPTS den in Abb. 21 wiedergegebenen Inhalt (wobei aus Einfachheitsgründen von den beiden Lösungskonzepten nur IDENTICAL.LETTERS definiert ist; die hier in Programmzeile 10 verwendete LOGO-Anweisung ‚local‘ dient zum Einrichten einer lokalen Variable im Arbeitsspeicher, die nur für die Laufzeit der Funktion Gültigkeit hat).

```
:LETTER.CONCEPTS: is
    „IDENTICAL.LETTERS ALPHABETICAL.SEQUENCE“

to IDENTICAL.LETTERS
10 local „WORD“
20 make „WORD“ :ITEM:
30 if CONTAINS butfirst of :WORD: first of :WORD:
    then make „IDENTICAL.LETTERS“
        sentence of „YES“ and first of :WORD: , stop
    else make „WORD“ butfirst of :WORD:
40 if (count of :WORD:) = 1
    then make „IDENTICAL.LETTERS“ „NO ONE“ , stop
    else go to line 30
end
```

Abb. 21: Inhalt der Datei LETTER CONCEPTS.

(2) Die LOGO-Operation ‚first of :LETTER.CONCEPTS:‘ im Aktionsteil von SILEX2 wird ermöglicht, nachdem mit SILEX1 die Datei LETTER CONCEPTS aktiviert worden ist.

(3) Die LOGO-Anweisung ‚do :CONCEPT:‘ im Aktionsteil von SILEX4 beinhaltet die Ausführung des Lösungskonzeptes als eine Programmfunktion, im vorliegenden Fall also die Ausführung von IDENTICAL.LETTERS (vgl. Abb. 21).

(4) Die Abfrage ‚first of thing of :CONCEPT:‘ im Konditionalteil von SILEX5 und SILEX6 bezieht sich auf das Ergebnis der Funktionsausführung des Lösungskonzeptes, im Beispiel also auf das Ergebnis der Programmfunktion IDENTICAL.LETTERS (vgl. deren Programmzeilen 30 bzw. 40 in Abb. 21).

(5) Das Abarbeiten der einzelnen Buchstabengruppen wird vom Benutzer durch die Eingabe des Klingesignals \*BELL (in SILEX3) abgeschlossen und im Konditionalteil von SILEX7 bzw. SILEX8 von dem Programm mit der Abfrage ‚ITEM: = :bell:‘ erkannt, worauf je nach dem Ergebnis von ‚(count of :UNSUITABLE.ITEM:) = 1‘ entweder SILEX7 oder SILEX8 „feuert“ (so der Ausdruck in der Terminologie von Produktionssystemen).

Damit die in Abb. 20 wiedergegebenen Produktionsregeln in der erwünschten Weise arbeiten können, müssen sie zu einem Produktionssystem zusammengefaßt werden, das vom Interpreter gelesen und ausgeführt werden kann. Die Definition eines Produktionssystems ist eine relativ einfache Aufgabe, wie aus der Darstellung in Abb. 22 zu ersehen ist: Die einzelnen Produktionsregeln werden lediglich auf einer Liste ihrem Namen nach in eine bestimmte Reihenfolge gebracht, die für die Abarbeitung durch den Interpreter von Bedeutung ist.

```
to SINGLE.LETTER.EXCLUSION
10 make „PRODUCTION.LIST“
    „SILEX5 SILEX6 SILEX7 SILEX8 SILEX1 SILEX2 SILEX3 SILEX4“
end
```

Abb. 22: Definition des Produktionssystems für die Aufgabe des „Unpassenden Streichens“.

Der Interpreter selbst ist ein einfaches sequentielles LOGO-Programm mit einigen Unterprogrammen, dargestellt in den Abb. 23 und 24.

Die Arbeitsweise des Interpreters läßt sich wie folgt beschreiben:

(1) Mit RUN :PRODUCTION.SYSTEM: wird der Programmablauf eines Produktionssystems gestartet, im vorliegenden Fall beispielsweise mit RUN „SINGLE.LETTER.EXCLUSION“.

(2) In Programmzeile 10 von RUN wird durch die ‚do‘-Anweisung die Funktionsdefinition des Produktionssystems ausgeführt, was nichts anderes besagt, als daß im Arbeitsspeicher die entsprechende Produktionsliste aktiviert wird (vgl. die Funktionsdefinition von SINGLE.LETTER.EXCLUSION in Abb. 22).



```

to RUN :PRODUCTION.SYSTEM:
10 do :PRODUCTION.SYSTEM:
20 PROCESS :PRODUCTION.LIST:
30 if ACTIVE :HALT:
    then stop
    else go to line 20
end

```

Abb. 23: Hauptprogramm des Interpreters für Produktionssysteme.

```

to PROCESS :PRODUCTIONS:
10 if :PRODUCTIONS: = :empty:
    then make „HALT“ „NO PRODUCTIONS READY“ , stop
20 if READY first of :PRODUCTIONS:
    then FIRE first of :PRODUCTIONS:
    else PROCESS butfirst of :PRODUCTIONS:
end

to READY :PRODUCTION:
10 do butfirst of text of :PRODUCTION: 10
20 iftrue output „true“
30 iffalse output „false“
end

to FIRE :PRODUCTION:
10 do butfirst of text of :PRODUCTION: 20
end

to ACTIVE :NAME:
10 if :NAME: = :empty:
    then output „false“
    else output „true“
end

to ,
end

```

Abb. 24: Unter- und Hilfsprogramme des Interpreters.

(3) Programmzeile 20 setzt den Prozeß des Abarbeitens der Produktionsliste in Gang; dieser Prozeß ist - wie aus dessen Funktionsdefinition in Abb. 24 hervorgeht - ein rekursiver Vorgang des Suchens nach der *ersten* ausführbaren Produktion (per READY getestet und per FIRE ausgeführt, wobei mittels

der LOGO-Operation ‚text‘ auf die jeweilige Programmzeile der entsprechenden Produktionsregel zugegriffen wird). Der Interpreter folgt dabei dem Dominanzprinzip der Regelabarbeitung, weshalb die Reihenfolge der Produktionsregeln in der Produktionsliste von Bedeutung ist, um Konflikte bei der Ausführung von Produktionen zu vermeiden, wenn zu einem bestimmten Zeitpunkt der Konditionalteil mehrerer Produktionsregeln gleichzeitig erfüllbar ist.

(4) In Programmzeile 30 wird schließlich geprüft, ob im Arbeitsspeicher das Haltsignal - mit welchem Wert auch immer - gesetzt ist, worauf im positiven Fall die Interpretation des Produktionssystems abgeschlossen ist („stop“), im negativen Fall jedoch so lange fortgesetzt wird, bis das Haltsignal erscheint (entweder über den Aktionsteil einer Produktionsregel wie in SILEX7 oder über Programmzeile 10 von PROCESS, wenn keine der Produktionen ausführbar ist).

Eine anschauliche Vorstellung von dem konkreten Ablauf der Informationsverarbeitung beim Lösen einer Aufgabe des „Unpassenden Streichens“ gewinnt man, wenn man sich einen Programmlauf per RUN „SINGLE.LETTER.EXCLUSION“ ansieht und mit den empirischen Daten vergleicht, wie dies in Tabelle 3 dargestellt ist. (Anmerkung: Am Terminal ist mit dem Programmaufruf RUN „SINGLE.LETTER.EXCLUSION“ tatsächlich nur das zu sehen, was in Tabelle 3 als [Eingabe] bzw. [Ausgabe] spezifiziert ist. Um einen vollständigen Überblick über die zum jeweiligen Zeitpunkt „feuernden“ Produktionen und die damit resultierenden Arbeitsspeichereinhalte zu bekommen, ist der Interpreter um entsprechende Druckanweisungen zu erweitern: In Programmzeile 20 von RUN ist an ‚PROCESS :PRODUCTION.LIST:‘ die LOGO-Anweisung ‚list all names‘ anzuhängen, die den gesamten Arbeitsspeichereinhalt am Terminal ausdruckt, und in Programmzeile 20 von PROCESS ist vor ‚FIRE first of :PRODUCTIONS:‘ die Druckanweisung ‚print first of :PRODUCTIONS:‘ zu schreiben, um sehen zu können, welche Produktion zu diesem Zeitpunkt feuert.)

### *3.2.3 Transparenz und Abbildtreue von Produktionssystemen*

Der Vergleich des in Tabelle 3 wiedergegebenen Programmlaufs des Produktionssystems SINGLE.LETTER.EXCLUSION mit den beigefügten Verbalisierungsdaten, aber auch deren in Tabelle 2 mitgeteilter Protokollanalyse zeigt einen Auflösungsgrad der abgebildeten Vorgänge des Problemlösens und einen Annäherungsgrad an die Realität menschlicher Informationsverarbeitung, die mit den Mitteln des herkömmlichen Programmierens, insbesondere des „top-down programming“, in dieser Transparenz und Abbildtreue kaum zu erreichen ist. Das läßt sich noch stärker verdeutlichen, wenn man statt des Produk-

Tabelle 3: Programmlauf des Produktionssystems SINGLE.LETTER.  
EXCLUSION.

PR	Arbeitsspeicher-Veränderungen durch Aktionen der Produktions- regeln (PR)	Protokollausschnitt
SILEX1	+AABC ACAD ACSH AACG [Eingabe] :GIVEN.ITEMS: is „AABC ACAD ACSH AACG“ :LETTER.CONCEPTS: is „IDENTICAL.LETTERS ALPHABETICAL.SEQUENCE“	„Also, wir haben vier Buchstaben- gruppen hier: AABC, ACAD, ACSH und AACG.“
SILEX2	:CONCEPT: is „IDENTICAL.LETTERS“	„Woll’n zuerst mal seh’n, welche von denen gleiche Buch- staben drin haben.“
SILEX3	*AABC [Eingabe] :ITEM: is „AABC“	„AABC“
SILEX4	:IDENTICAL.LETTERS: is „YES A“	„hat zwei A’s,”
SILEX5	:IDENTICAL.LETTERS: is „“ :ITEM: is „“	
SILEX3	*ACAD [Eingabe] :ITEM: is „ACAD“	„ACAD“
SILEX4	:IDENTICAL.LETTERS: is „YES A“	„auch,“
SILEX5	:IDENTICAL.LETTERS: is „“ :ITEM: is „“	
SILEX3	*ACSH [Eingabe] :ITEM: is „ACSH“	„ACSH“
SILEX4	:IDENTICAL.LETTERS: is „NO ONE“	„nicht;“
SILEX6	:UNSUITABLE.ITEM: is „ACSH“ :IDENTICAL.LETTERS: is „“ :ITEM: is „“	„aha!“
SILEX3	*AACG [Eingabe] :ITEM: is „AACG“	„Aber AACG“

Tabelle 3 : Fortsetzung

SILEX4	:IDENTICAL.LETTERS: is „YES A“	„wieder.“
SILEX5	:IDENTICAL.LETTERS: is „, :ITEM: is „,	
SILEX3	* BELL :ITEM: is :bell:	[Eingabe]
SILEX7	ACSH :HALT: is „PROBLEM IS SOLVED“	[Ausgabe] „Also ist ACSH die unpassende Gruppe, paßt nicht zu den ändern.“

tionssystem ein sequentielles Programm für das „Single Letter Exclusion“ schreibt, wie es in Abb. 25 wiedergegeben ist.

In seinem Verhalten ist das sequentielle Programm zwar identisch mit dem Produktionssystem, die Transparenz der Informationsverarbeitung, insbeson-

```

to SINGLE.LETTER.EXCLUSION
10 make „GIVEN.ITEMS“ request
20 get LETTER CONCEPTS
30 make „CONCEPT“ first of :LETTER.CONCEPTS:
40 make „ITEM“ request
50 if both (count of :UNSUITABLE.ITEM:) = 1 and :ITEM: = :bell:
    then print :UNSUITABLE.ITEM: , stop
60 if both not (count of :UNSUITABLE.ITEM:) = 1 and :ITEM: = :bell:
    then make „LETTER.CONCEPTS“
        REMOVE :CONCEPT: from :LETTER.CONCEPTS: ,
        go to line 30
70 do :CONCEPT:
80 if (first of thing of :CONCEPT:) = „NO“
    then make „UNSUITABLE.ITEM“
        sentence of :ITEM: and :UNSUITABLE.ITEM:
90 go to line 40
end

```

Abb. 25: Sequentielles LOGO-Programm für die Aufgabe des „Unpassenden Streichens“.

dere hinsichtlich der konkreten Arbeitsspeicher-Veränderungen, geht jedoch weitgehend verloren. Die im Programmcode noch sichtbaren Entsprechungen zwischen den Programmzeilen von Abb. 25 und den Produktionen von Abb. 20 sind im einzelnen:

- Programmzeile 10 und 20: SILEX1
- Programmzeile 30: SILEX2
- Programmzeile 40: SILEX3
- Programmzeile 50: SILEX7
- Programmzeile 60: SILEX8
- Programmzeile 70: SILEX4
- Programmzeile 80: SILEX6

Für Programmzeile 90 gibt es keine Entsprechung, ferner fällt Produktion SILEX5 ganz aus dem sequentiellen Programm heraus. Vor allem aber ist in dem Programm nicht mehr immer klar, unter welchen „Datenbedingungen“ - den im Konditionalteil von Produktionsregeln angesprochenen Arbeitsspeicherinhalten - einzelne Programmoperationen ausgeführt werden (vgl. Programmzeilen 10, 20, 30, 40 und 70), so daß die mit einem Programmlauf einhergehenden Arbeitsspeicher-Veränderungen nicht mehr so leicht nachvollziehbar sind wie bei der Abarbeitung des Produktionssystems.

Noch entscheidender sind die Mängel einer sequentiellen Programmierung hinsichtlich der Abbildtreue eines so erstellten Simulationsmodells. Die Modellarchitektur von Produktionssystemen - Arbeitsspeicher als Kurzzeitgedächtnis, Produktionsspeicher als Langzeitgedächtnis und Interpreter als „kognitive Exekutive“ - ist konstitutiv für die psychologische Relevanz eines Simulationsmodells, im Falle von sequentiellen Programmen wird aber beispielsweise über Arbeitsspeicher und Interpreter überhaupt nichts ausgesagt, obwohl auch hier für den konkreten Programmlauf beide Komponenten gegeben sind (nämlich in Form des von der verwendeten Programmiersprache zur Verfügung stehenden Arbeitsspeichers und Interpreters). Erst die Einbeziehung einer umfassenden Modellarchitektur wie die mit der Produktionssystem-Konzeption verbundenen macht es möglich, über Transparenz und Abbildtreue von Simulationsmodellen sinnvoll diskutieren zu können.

Als bemerkenswertestes Beispiel einer mit der Produktionssystem-Konzeption verknüpften psychologischen Theorienbildung sei nur auf die ACT-Theorie von Anderson (1976) verwiesen, in der der bis heute weitgehendste Versuch unternommen wurde, Strukturmodelle des menschlichen Gedächtnisses - mittels propositionaler semantischer Netze („deklaratives Wissen“ oder „Wissen Was“) - mit Prozeßmodellen der Informationsverarbeitung - mittels informationeller Produktionssysteme („prozedurales Wissen“ oder „Wissen Wie“) - zu einer einheitlichen Theorie kognitiver Aktivität im Bereich des Sprachverstehens und der Sprachproduktion zu verbinden. Trotz aller Kritik

(vgl. z.B. Wexler, 1978) verdient diese Theorie nicht nur ihres psychologischen Anspruchs, sondern auch ihres empirischen Gehaltes wegen Aufmerksamkeit, denn die Kriterien der psychologischen Relevanz und der empirischen Testbarkeit sind gerade im Zusammenhang mit der Computer-Simulation besonders schwer zu beurteilen (vgl. Abschnitt 4, Validierung und Anwendbarkeit von Simulationsmodellen).

### 3.3 Das Interpreterproblem von Produktionssystemen

Obwohl die Produktionssystem-Konzeption als die im Zusammenhang mit der Simulationsmethodik bisher am weitesten entwickelte psychologische Theorienbildung gelten kann, ist das Interpreterproblem, die Frage nach der Bedeutsamkeit des Interpreters als der „zentrale Prozessor“ oder die „kognitive Exekutive“ der menschlichen Informationsverarbeitung, erst in Ansätzen diskutiert worden (vgl. Ueckert, 1980b). Welchen Anforderungen ein effizienter Interpreter unter anderem genügen sollte, haben beispielsweise McDermott & Forgy (1978) ausgeführt:

- Der Interpreter sollte das informationsverarbeitende System, während es an einer bestimmten Aufgabe arbeitet, in allen Versuchen unterstützen, sensitiv für die Vielfältigkeit seiner externen Informationsquellen zu bleiben.
- Ebenso sehr sollte der Interpreter das System in die Lage versetzen, sensitiv für seine eigenen, internen Tätigkeiten zu sein.
- Der Interpreter sollte fähig sein, mit „widersprüchlichen Daten“ umgehen zu können, insbesondere zwischen zu einem Zeitpunkt relevanter und nicht mehr relevanter (oder noch nicht relevanter) Information unterscheiden können.
- Hierbei sollte der Interpreter vor allem erkennen können, ob sich aufgrund eines gegebenen Arbeitsspeicherinhalts gleichzeitig mehrere Produktionen aktivieren lassen, und gegebenenfalls geeignete „Konfliktlösungsmöglichkeiten“ anbieten können.

Konkret formuliert kann das Interpreterproblem unter folgenden Fragestellungen behandelt werden:

- (1) Welche Lesarten von Produktionsregeln sind von dem Interpreter realisierbar?
- (2) Welche Möglichkeiten der *Konfliktlösung* gibt es für den Interpreter bei gleichzeitiger Erfüllbarkeit mehrerer Konditionalteile von Produktionsregeln?

- (3) Welche Formen von *Lernfähigkeit* muß der Interpret für eine realitäts-gerechte Informationsverarbeitung aufweisen?
- (4) Welche „*Bewußtseinsfunktionen*“ - wenn überhaupt - sollte der Interpret als „kognitive Exekutive“ eines informationsverarbeitenden Systems ausführen können?

Eine einführende Diskussion dieser Fragestellungen soll in den folgenden Unterabschnitten gegeben werden.

### 3.3.1 Lesarten von Produktionsregeln

In seiner bisher beschriebenen Funktionsweise arbeitet der Interpret alle Produktionsregeln von links nach rechts ab (vgl. die Definition von PROCESS in Abb. 24): Er liest zuerst den linken Teil einer Produktion, den Konditionalteil, und prüft, ob dieser im Arbeitsspeicher erfüllt ist (mittels ‚do butfirst of text of :PRODUCTION: 10‘ in READY); ist dies der Fall, dann führt er den rechten Teil der Produktion aus, den Aktionsteil (mittels ‚do butfirst of text of :PRODUCTION: 20‘ in FIRE). Diese Arbeitsweise von links nach rechts wird als *datengesteuert* („data-driven“, „condition-driven“) bezeichnet, da die Daten des Arbeitsspeichers - die augenblicklichen Arbeitsspeicherinhalte - die aktuelle Interpretation des Produktionssystems bestimmen.

Es ist jedoch auch die umgekehrte Lesart realisierbar, ein Abarbeiten der Produktionsregeln von rechts nach links: Der Interpret liest zuerst den rechten Aktionsteil einer Produktion und prüft - um die in diesem angegebenen Operationen ausführen zu können -, ob die im linken Konditionalteil dieser Produktion spezifizierten Daten im Arbeitsspeicher gegeben sind; ist dies der Fall, kann der Aktionsteil ausgeführt werden, ist dies jedoch nicht der Fall, sucht sich der Interpret diejenige Produktionsregel, in deren Aktionsteil die Daten des ersteren, noch nicht erfüllbaren Konditionalteils erzeugt werden. Dieser Prozeß wird so lange fortgesetzt, bis ein im Arbeitsspeicher erfüllbarer Konditionalteil gefunden ist und dessen zugehöriger Aktionsteil ausgeführt werden kann. Diese Arbeitsweise des Interpreters wird als *handlungsgesteuert* („action-driven“, „goal-driven“) bezeichnet, da die angestrebten Handlungen - die Aktionen von Produktionen - die Interpretation des Produktionssystems bestimmen.

Auf eine Kurzformel gebracht, lassen sich die beiden Lesarten von Produktionsregeln so beschreiben:

- (1) Datengesteuerte Interpretation:  
für alle N: wenn K erfüllt ist, tue A.
- (2) Handlungsgesteuerte Interpretation:  
für alle N: um A tun zu können, erfülle K.

(Wobei N den Namen, K den Konditionalteil und A den Aktionsteil einer Produktion bezeichnet.)

Die psychologische Bedeutsamkeit der beiden unterschiedlichen Interpretationsweisen liegt auf der Hand. Datengesteuerte Interpretation ist überall da angebracht, wo es auf schnelle, angepaßte, situationsgerechte Informationsverarbeitung ankommt, insbesondere mit gut ausgearbeiteten oder „überlerten“ Programmen. Handlungsgesteuerte Interpretation ist dagegen „offener“; man kann sie als eine Art von „Probearbeiten“ auffassen, als Denken im engeren Sinne, das beim Planen, Beweisen, Schlußfolgern, aber konkret auch beim Ausarbeiten und Verbessern von kognitiven Produktionssystemen eine Rolle spielt. Als Beispiel hierzu könnte man sich einen handlungsgesteuerten Lauf des Produktionssystems SINGLE.LETTER.EXCLUSION vorstellen, bei dem im Arbeitsspeicher lediglich der Ausdruck `„:UNSUITABLE.ITEM: is „ACSH“` gegeben ist, von dem zu beweisen sei, daß er die richtige Lösung einer Aufgabe des „Unpassenden Streichens“ am Beispiel von „AABC ACAD ACSH AACG“ darstellt. Der Leser möge sich selbst vergegenwärtigen, welche Produktionen der Interpreter hier in welcher Reihenfolge zu betrachten (und gegebenenfalls auch zu feuern) hätte, um einen folgerichtigen Beweis der Richtigkeit von `„:UNSUITABLE.ITEM: is „ACSH“` für `„:GIVEN.ITEMS: is „AABC ACAD ACSH AACG“` vorzulegen.

Offensichtlich müßte für die handlungsgesteuerte Arbeitsweise das Hauptprogramm des Interpreters in Abb. 23 um eine entsprechend zu definierende Prozeßfunktion - etwa als TRY.PROCESS zu bezeichnen - verändert oder ergänzt werden (unter Berücksichtigung von Bedingungen, wann PROCESS und wann TRY.PROCESS aktiviert werden sollen).

### 3.3.2 Konfliktlösungsstrategien („*conflict resolution*“)

In Anbetracht der Tatsache, daß jede Produktionsregel eine in sich geschlossene Einheit darstellt („Modularität“), ist es insbesondere bei umfangreicheren Produktionssystemen möglich, daß aufgrund des Arbeitsspeicherinhalts mehrere Produktionen gleichzeitig - und das zu unterschiedlichen Zeiten immer wieder - feuern könnten. Die Theorie der Produktionssysteme verlangt jedoch, daß zu einem Zeitpunkt stets nur genau *eine* Produktion ausgeführt werden kann, wie klein oder groß auch immer der Zeittakt sein mag. Das Problem ist also, welche der Produktionen der Interpreter dann aktivieren soll.

In der Literatur sind bisher recht unterschiedliche Konfliktlösungsmöglichkeiten für diesen Fall vorgeschlagen worden. McDermott & Forgy (1978) beispielsweise diskutieren die folgenden Möglichkeiten:



- (1) Reihungs-Dominanz der Produktionsregeln: Die erste Produktion innerhalb des Produktionssystems, deren Konditionalteil erfüllt ist, wird ausgeführt. (In diesem Sinne arbeitet der in Abb. 23 und 24 wiedergegebene Interpreter.)
- (2) Spezialfall-Dominanz: Produktionen mit einem spezifischen Konditionalteil werden allgemeineren Produktionen vorgezogen. (Von den Produktionsregeln in Abb. 20 ist beispielsweise SILEX3 ein Spezialfall von SILEX4.)
- (3) Neuheits-Dominanz („recency“): Zuletzt gefeuerte Produktionen oder solche, die zuletzt erfüllte Datenelemente in ihrem Konditionalteil aufweisen, werden bevorzugt.
- (4) Unterschiedlichkeits-Dominanz („distinctiveness“): Möglichst in ihrem Konditionalteil und/oder Aktionsteil von vorangehenden Produktionen verschiedene Regeln werden vorgezogen.
- (5) Zufallsauswahl in Ermangelung anderer Kriterien.

Abgesehen von der Reihungs-Dominanz und der Zufallsauswahl sind diese - und weitere, von anderen Autoren (z.B. Davis & King, 1977; Rychener & Newell, 1978) zitierte - Konfliktlösungsmöglichkeiten nicht immer eindeutig, d.h. sie führen in vielen Fällen nicht zu genau einer ausführbaren Produktion. Sinnvoll ist daher die geeignete Kombination dieser - in sich auch noch weiter unterteilbarer - Möglichkeiten zu ausgefeilten „Konfliktlösungsstrategien“, wie sie von McDermott & Forgy diskutiert werden, insbesondere auch unter dem Aspekt, in welcher Weise sie die eingangs erwähnten Anforderungen an einen effizienten Interpreter zu stützen erlauben.

### 3.3.3 *Adaptivität (Lernfähigkeit) von Produktionssystemen*

Eine der wichtigsten Aufgaben des Interpreters ist die Fähigkeit, Produktionssysteme für sich ändernde Anforderungen an eine situationsgerechte Informationsverarbeitung adaptiv zu halten. Theoretisch ist das Problem der Adaptivität oder Lernfähigkeit von Produktionssystemen leicht zu lösen: Produktionsregeln können jederzeit in einem Produktionssystem

- hinzugefügt,
- entfernt,
- generalisiert,
- spezialisiert

werden. Die Frage der praktischen Realisierbarkeit derartiger Modifikationsmöglichkeiten ist ein empirisches - oder auch nur technisches - Problem:

- Lernen durch Belehrung,
- Lernen durch Beispiele,

- Erfolgskontrolle durch Rückmeldung,
- Generalisieren durch Meta-Regeln

sind einige der in der Literatur bisher behandelten Formen der Adaptivität von Produktionssystemen (vgl. beispielsweise das Kapitel „Learning“ in Waterman & Hayes-Roth, 1978).

Ein konkretes Beispiel zur Diskussion des Adaptivitätsproblems wird in Abschnitt 3.4 vorgestellt.

### 3.3.4 „Bewußtseinsfunktionen“ des Interpreters

Die Frage, inwieweit der Interpreter von Produktionssystemen „Bewußtseinsfunktionen“ ausüben sollte, wie sie für den Bereich menschlicher Kognition charakteristisch sind, ist ein bisher noch kaum diskutiertes Problem der Simulationsmethodik. Unter dem Aspekt des „zentralen Prozessors“ oder der „kognitiven Exekutive“ von informationsverarbeitenden Systemen kann diese Frage jedoch nicht ausgeklammert werden, und die Produktionssystem-Konzeption bietet mit ihrer Modellarchitektur den bisher brauchbarsten Ansatz zu deren Diskussion.

Nach einem Vorschlag von Ueckert (1980b) kann man zwei Grundfunktionen des menschlichen Bewußtseins unterscheiden, deren Realisierbarkeit durch den Interpreter als „kognitive Exekutive“ von Produktionssystemen gegeben erscheint:

- (1) Eine *Zeigerfunktion* derart, daß der Interpreter jederzeit in der Lage ist, auf den Inhalt eines beliebigen Arbeitsspeichers zu „zeigen“, was im Bereich des menschlichen Bewußtseins der Fähigkeit der Aufmerksamkeitslenkung auf beliebige Bewußtseinsinhalte entspricht.
- (2) Eine *Übersetzerfunktion* dahingehend, daß der Interpreter fähig ist, beliebige Arbeitsspeicherinhalte mit Hilfe geeigneter Operationen nach außen zu „übersetzen“, d.h. im Sinne der menschlichen Fähigkeit zur sprachlichen und/oder manuellen Umsetzung von Bewußtseinsinhalten alles das zu externalisieren, was durch die Zeigerfunktion zu einem Zeitpunkt im Fokus der Aufmerksamkeit gehalten werden kann.

Es ist naheliegend, daß ein effizienter Interpreter, der in gleicher Weise eine datenorientierte wie handlungsorientierte Informationsverarbeitung realisieren, Konfliktlösungsmöglichkeiten für konkurrierende Produktionen bereitstellen und Formen einer situationsgerechten Adaptivität aufweisen soll, über beide „Bewußtseinsfunktionen“ verfügen muß: Die Zeigerfunktion ermöglicht die unterschiedlichen Lesarten von Produktionsregeln und die Entwicklung

von Strategien der Konfliktlösung, die Übersetzerfunktion ist Voraussetzung für die Anpassungsleistungen des informationsverarbeitenden Systems in seiner Auseinandersetzung mit der Umwelt. Inwiefern diese Funktionen jedoch heute schon in einem einzigen Interpreter implementiert - d.h. in Form eines Computer-Modells programmiert - werden können, ist noch eine offene Frage; würde ihre Beantwortung doch bedeuten, auch von einem „Bewußtsein der Maschinen“ sprechen zu können.

### 3.4 „Künstliche Intelligenz“ oder:

Wie man dem Rechner das Rechnen beibringen kann

Im Grunde genommen ist jedes Simulationsmodell, das Prozesse der Informationsverarbeitung auf dem Rechner nachbildet, eine Form von „künstlicher Intelligenz“ (KI), da die Realisierung des Modells auf einem artifiziellen System - dem Computer - erfolgt und die zugrundeliegenden Prozesse informationell - und damit, wenn man so will, „geistig“ oder „intelligent“ - sind. Von diesem weiteren Begriff der „künstlichen Intelligenz“ abzuheben ist ein engerer Begriff, der für die Entwicklung künstlicher Systeme im Bereich der Informatik charakteristisch ist: „Künstliche Intelligenz“ weist jedes Computer-Programm auf, das *Leistungen* produziert, die, wenn beim Menschen beobachtet, als „intelligent“ zu bezeichnen wären, wobei es keine Rolle spielt, ob die den Leistungen zugrundeliegenden *Vorgänge* „menschenähnlich“ sind oder nicht. Beispiele für diese Form „künstlicher Intelligenz“ gibt es zu den unterschiedlichsten Leistungsbereichen, wie jedes Buch zu diesem Forschungsgebiet belegt (als lesenswerte allgemeinverständliche Einführung vgl. das Buch von Boden, 1977).

Inzwischen hat sich auch auf diesem Gebiet die Produktionssystem-Konzeption so weit durchgesetzt, daß heute kaum noch ein neues KI-System geschrieben wird, ohne auf diesen Ansatz zu rekurrieren (vgl. Waterman & Hayes-Roth, 1978). Insbesondere die mit Produktionssystemen verbundene Flexibilität und Adaptivität hat sich als entscheidender Vorteil nicht nur in der Entwicklung von Simulationsmodellen, sondern auch in der Konstruktion von KI-Systemen herausgestellt. Dies mag an einem simplen Beispiel demonstriert werden: einem „lernfähigen“ Produktionssystem zum Addieren zweier ganzer positiver Zahlen.

Das Modell, das zunächst nicht als ein Simulationsmodell konzipiert ist - und von daher als ein KI-System im engeren Sinne zu verstehen ist -, soll die Addition auf die elementaren Operationen des Zählens (in der Programmiersprache LOGO mit der ‚count‘-Operation gegeben) und der Konkatenation (Zusammenfügen zweier Einheiten zu einer neuen Einheit; in LOGO mit der ‚word‘- bzw. der ‚sentence‘-Operation ausführbar) zurückführen. Die Adap-

tivität oder „Lernfähigkeit“ des Modells soll so realisiert werden, daß das Ergebnis einer Addition mit zwei vorgegebenen Zahlenwerten in einer „Additionstabelle“ langfristig gespeichert wird, so daß bei einer erneuten Vorgabe der beiden Zahlenwerte das Ergebnis nur noch aufgesucht und ausgegeben werden braucht. Die Flexibilität, die sich in der Produktionssystem-Konzeption dadurch ergibt, daß sich Produktionssysteme wechselseitig aufrufen können, wird für dieses KI-System auf den einfachsten Fall reduziert, den wechselseitigen Aufruf zweier Produktionssysteme: (1) ADD, in dem das eigentliche „Rechnen“ (durch Zählen und Konkatenation) stattfindet, und (2) ADD.TABLE, in dem die „Rechenergebnisse“ nach und nach gespeichert werden. Wie die Interaktion dieser beiden Produktionssysteme bewerkstelligt wird, ist aus der - wiederum im LOGO-Formalismus gehaltenen - Darstellung in den Abb. 26 und 27 zu ersehen.

Der wechselseitige Aufruf dieser beiden Produktionssysteme erfolgt in den Produktionsregeln ADD1, TAB1 und ADD5, und zwar, wie ersichtlich, auf unterschiedliche Weise: In ADD1 und TAB1 wird das jeweils andere Produktionssystem durch die ‚do‘-Anweisung lediglich über eine Veränderung der

```

to ADD
  10 make „PRODUCTION.LIST“ „ADD1 ADD2 ADD3 ADD4 ADD5“
end

ADD1
both :NUMBER1: = :empty: and :NUMBER2: = :empty:
→ make „NUMBER1“ request ,
   make „NUMBER2“ request ,
   do „ADD.TABLE“

ADD2
not (count of :TALLY1 :) = :NUMBER1:
→ make „TALLY1“ word of „X“ and :TALLY1:

ADD3
not (count of :TALLY2) = :NUMBER2:
→ make „TALLY2“ word of „X“ and :TALLY2:

ADD4
:SUM: = :empty:
→ make „SUM“ count of word of :TALLY1: and :TALLY2:

ADD5
ACTIVE :SUM:
→ make „PRODUCTION.SYSTEM“ „ADD.TABLE“,
   do :PRODUCTION.SYSTEM:

```

Abb. 26: Produktionssystem für das Addieren zweier ganzer positiver Zahlen.

```

to ADD.TABLE
10 make „PRODUCTION.LIST“ „TAB1 TAB2“
end

TAB1
:SUM: = :empty:
→ do „ADD“

TAB2
ACTIVE :SUM:
→ make „NEW.PRODUCTION“
    (words „TAB.“ :NUMBER1: „“ :NUMBER2:) ,
    make „NEW.CONDITION“
        (sentences „both :NUMBER1 : = “ :NUMBER1 :
            „and :NUMBER2: =“ :NUMBER2:) ,
    make „NEW.ACTION“
        (sentences „print“ :SUM: „“
            „make“ :quote: „HALT“ :quote:
            :quote: „THE SUM HAS BEEN FOUND“ :quote:) ,
    make „HALT“ „THE SUM HAS BEEN COMPUTED AND
        ENTERED INTO THE TABLE“ ,
    print :SUM:

```

Abb. 27: Produktionssystem zum Erzeugen einer Additionstabelle.

*Liste* der Produktionen, nicht jedoch auch des *Namens* des Produktionssystems aktiviert (was einer hierarchischen - oder übergeordneten - Abhängigkeit der Produktionssysteme entspricht); in ADD5 dagegen wird auch der Name des Produktionssystems verändert und der Sprung per ‚do :PRODUCTION.SYSTEM:‘ ausgeführt (dies entspricht eher einer heterarchischen - oder nebengeordneten - Abhängigkeit). Der unterschiedliche Gebrauch ergibt sich aus der Zielsetzung für die Interaktion der beiden Produktionssysteme. In ADD wird mit der Produktion ADD1 zunächst einmal die Eingabe der beiden Zahlenwerte abgefragt, worauf mit einem Sprung nach ADD.TABLE festzustellen ist, ob das Ergebnis dort schon gespeichert ist; anfangs ist das natürlich noch nicht der Fall, weshalb über TAB1 der Rücksprung nach ADD erfolgt. Sodann wird über die Produktionen ADD2, ADD3 und ADD4 das Berechnen des Ergebnisses durch zwei „X-Strichlisten“ (in ADD2 bzw. ADD3) und das Auszählen der Länge der daraus gebildeten Gesamtliste (in ADD4) durchgeführt. Damit ist die Summe berechnet und es erfolgt über ADD5 ein Sprung in das Produktionssystem ADD.TABLE, um in diesem das Ergebnis abzuspeichern, was mittels Produktion TAB2 vorbereitet wird: Im Arbeitsspeicher werden der Name für eine neue Produktionsregel sowie ein neuer Konditionalteil und ein neuer Aktionsteil eingerichtet (dabei sind die LOGO-Operatio-

nen ‚words‘ und ‚sentences‘ Erweiterungen der einfachen Operationen ‚word‘ und ‚sentence‘ - zum Zusammensetzen eines LOGO-Wortes bzw. eines LOGO-Satzes - mit beliebig vielen Argumenten; die LOGO-Variable :quote: hat das Anführungszeichen „ als ihren Wert). Der Sinn der Aktionen in TAB2 ist, das berechnete Additionsergebnis als eine neue Produktionsregel in die Produktionsliste von ADD.TABLE aufzunehmen.

Um diese primitive Form von „Lernfähigkeit“ realisieren zu können, muß der Interpreter um diese Möglichkeit der Adaptivität erweitert werden: Er muß, noch bevor er das Haltsignal im Arbeitsspeicher liest, auf :NEW.PRODUCTION: mit dem Schreiben einer neuen Produktionsregel (in LOGO also dem Schreiben einer neuen Funktionsdefinition) reagieren können. Der bereits in Abb. 23 (Abschnitt 3.2.2) vorgestellte Produktionssystem-Interpreter ist nunmehr um die Programmzeile

```
25 if ACTIVE :NEW.PRODUCTION:
    then EXPAND :PRODUCTION.SYSTEM:
```

zu ergänzen. Das hierin verwendete Unterprogramm EXPAND ist in Abb. 28 definiert.

```
to EXPAND :PRODUCTION.SYSTEM:
10 do sentence „to“ :NEW.PRODUCTION:
15 do sentence „10 test“ :NEW.CONDITION:
20 do sentence „20 iftrue“ :NEW.ACTION:
25 do „end“
30 do sentence „erase“ :PRODUCTION.SYSTEM:
35 do sentence „to“ :PRODUCTION.SYSTEM:
40 do (sentences
    „10 make“ :quote: „PRODUCTION.LIST“ :quote:
    :quote: :NEW.PRODUCTION: :PRODUCTION.LIST: :quote:)
45 do „end“
50 make „NEW.PRODUCTION“ :empty:
55 make „NEW.CONDITION“ :empty:
60 make „NEW.ACTION“ :empty:
65 do :PRODUCTION.SYSTEM:
end
```

Abb. 28: Unterprogramm zum Erweitern eines Produktionssystems.

Dieses Unterprogramm ist im Grunde ein „Programm zum Schreiben von Programmen“: In den Programmzeilen 10-25 wird die Funktionsdefinition einer neuen Produktionsregel durchgeführt und in den Programmzeilen 35-45 die Funktionsdefinition des modifizierten, um die neue Produktionsregel erweiterten Produktionssystems, nachdem dessen alte Version mit der LOGO-Anweisung ‚erase‘ in Programmzeile 30 erst einmal gelöscht wurde.

Der Rest der EXPAND-Anweisung dient nur noch dem Löschen der nicht mehr benötigten Arbeitsspeicherinhalte (Programmzeilen 50-60) und dem Aktivieren des nunmehr veränderten Produktionssystems (Programmzeile 65). Aus der Funktionsdefinition von EXPAND wird nun verständlich, weshalb in Produktion ADD5 auch der Name des Produktionssystems (und nicht nur die Produktionsliste wie in ADD1 bzw. TAB1) geändert werden mußte: EXPAND erwartet den Wert der Variable :PRODUCTION.SYSTEM: als Argument (und das ist der Name eines Produktionssystems), um eben dieses Produktionssystem „expandieren“ zu können (im vorliegenden Fall also „ADD.TABLE“ und nicht „ADD“, von wo aus - in ADD5 - der Sprung erfolgt).

Der in Tabelle 4 dargestellte Programmlauf gibt einen Überblick über das Interaktionsgeschehen zwischen den beiden Produktionssystemen ADD und ADD.TABLE. Am Ende des ersten Programmlaufs hat der Interpreter die neue Produktionsregel

```

to TAB.3.2
10 test both :NUMBER1: = 3 and :NUMBER2: = 2
20 iftrue print 5,
    make „HALT“ „THE SUM HAS BEEN FOUND“
end

```

geschrieben und an den Anfang der Produktionsliste von ADD.TABLE gesetzt, so daß diese Produktion bei einem erneuten Programmlauf mit den gleichen Zahlenwerten das Ergebnis sofort ausgeben kann, ohne es nochmals berechnen zu müssen (vgl. den zweiten Programmlauf im zweiten Teil von Tabelle 4).

Inwieweit dieses KI-System auch als ein Simulationsmodell für das Addieren angesehen werden kann, bliebe zu diskutieren. Zumindest für die anfänglichen Fähigkeiten eines Kindes im Zahlenrechnen bietet es eine brauchbare Beschreibung: Das Aufstellen der beiden „X-Strichlisten“ ist recht ähnlich dem kindlichen „Fingerrechnen“, insbesondere bei einstelligen Zahlen, und auch das Behalten derartiger Rechenergebnisse -wenn auch vielleicht erst nach längerer Übung und nicht beim erstenmal wie in ADD.TABLE - erscheint kindgemäß. Von weit größerer Bedeutung für die „künstliche Intelligenz“-Forschung im engeren Sinne ist jedoch die Frage nach der Effizienz dieser „Addiermaschine“. Man kann sich leicht vorstellen, daß die Arbeitsweise mit ADD und ADD.TABLE schnell unökonomisch wird, wenn ADD mit großen Zahlen rechnen soll (was lange „X-Strichlisten“ ergäbe) oder wenn in ADD.TABLE umfangreiche Mengen von Rechenergebnissen zu speichern sind (was ein u.U. langwieriges Suchen nach einem bestimmten Ergebnis bedeutete). Beide Produktionssysteme müßten um geeignete Produktionen ergänzt werden, um die Arbeitsweise zu optimieren. Beispielsweise könnte auf die Zerleg-

Tabelle 4: Zwei Programmläufe der Produktionssysteme ADD und ADD.TABLE.

1. Lauf mit RUN „ADD“

PR	Arbeitsspeicher-Veränderungen
Start	:PRODUCTION.SYSTEM: is „ADD“ :PRODUCTION.LIST: is „ADD1 ADD2 ADD3 ADD4 ADD5“
ADD1	*3 [Eingabe] *2 [Eingabe] :NUMBER1: is „3“ :NUMBER2: is „2“ :PRODUCTION.LIST: is „TAB1 TAB2,,
TAB1	:PRODUCTION.LIST: is „ADD1 ADD2 ADD3 ADD4 ADD5“
ADD2	:TALLY1: is „X“
ADD2	:TALLY1: is „XX“
ADD2	:TALLY1: is „XXX“
ADD3	:TALLY2: is „X“
ADD3	:TALLY2: is „XX“
ADD4	:SUM: is „5“
ADD5	:PRODUCTION.SYSTEM: is „ADD.TABLE“ :PRODUCTION.LIST: is „TAB1 TAB2“
TAB2	:NEW.PRODUCTION: is „TAB.3.2“ :NEW.CONDITION: is „both :NUMBER1: = 3 and :NUMBER2: = 2“ :NEW.ACTION: is „print 5 , make „HALT“ „THE SUM HAS BEEN FOUND“ “ :HALT: is „THE SUM HAS BEEN COMPUTED AND ENTERED INTO THE TABLE“ 5 [Ausgabe]
Ende (nach EXPAND)	:PRODUCTION.LIST: is „TAB.3.2 TAB1 TAB2“ :NEW.PRODUCTION: is „ “ :NEW.CONDITION: is „ “ :NEW.ACTION: is „ “



Tabelle 4: Fortsetzung

2. Lauf mit RUN „ADD“ und der gleichen Eingabe

PR	Arbeitsspeicher-Veränderungen	
Start	:PRODUCTION.SYSTEM: is „ADD“ :PRODUCTION.LIST: is „ADD1 ADD2 ADD3 ADD4 ADD5“	
ADD1	+3 +2 :NUMBER1: is „3“ :NUMBER2: is „2“ :PRODUCTION.LIST: is „TAB.3.2 TAB1 TAB2“	[Eingabe] [Eingabe]
TAB.3.2	:HALT: is „THE SUM HAS BEEN FOUND“ 5	[Ausgabe]

barkeit von Zahlen im dekadischen System zurückgegriffen werden, um die Zähloperationen in ADD zu vereinfachen, oder es könnte die Kommutativität der Addition ausgenutzt werden, um den Umfang von ADD.TABLE zu reduzieren. Die Frage ist nur, *wer* diese weitergehende Form von Adaptivität - nämlich „Lernen aus der Unzulänglichkeit des Systemverhaltens“ - bewerkstelligen soll, der Modellkonstrukteur oder aber der Produktionssystem-Interpreter selbst, dem weitergehende Möglichkeiten des „Lernens“ als die simple EXPAND-Anweisung eingebaut sein sollten. Insbesondere wären hier - nach einiger Laufzeit von ADD (z.B. mit großen Zahlen) und ADD.TABLE (z.B. nach einer umfangreichen Tabellierung) - Phasen einer handlungsgesteuerten Interpretation (vgl. Abschnitt 3.3.1) sinnvoll, um zu einer Effektivitätsbeurteilung beider Produktionssysteme durch den Interpreter selbst gelangen zu können.

Eines der bemerkenswertesten KI-Systeme, das in dieser Richtung auf der Grundlage der Produktionssystem-Konzeption entwickelt wurde, ist das „AM-System“ von Lenat (1978, 1979). Das System „entdeckt“ mit einem Repertoire elementarer mathematischer Begriffe aus der Mengenlehre neue mathematische Konzepte und Relationen (wie z.B. Zahlbegriff, arithmetische Operationen, Primzahlpaare, Goldbachsche Vermutung, Diophantische Gleichungen, aber auch neuartige, in der Mathematik bisher *unbekannte* Begriffe wie z.B. „Zahlen mit maximal vielen Teilern“). Im Zahlenspiegel seiner Statistiken ist das AM-System schon ein recht interessantes „mathematisches Spielzeug“ : In etwa 1 Stunde Rechenzeit rekonstruiert es, wenn man so will, 100 Jahre Mathematikgeschichte auf der Basis von 115 Grundbegriffen und

250 Produktionsregeln und entwickelt 185 sehr differenzierte neue mathematische Konzepte (davon vom Autor 25 als „Gewinner“ und 60 als „Verlierer“ - neben 100 akzeptablen Begriffen - klassifiziert).

Dabei ist die Leistungsfähigkeit dieses KI-Systems nicht nur durch die Effektivität der 250 Produktionsregeln, die die „heuristische Suche“ in einem so großen Problemraum wie dem der Mathematik als einen „regelgeleiteten Explorationsprozeß“ gestalten helfen, sondern ebenso sehr durch die Arbeitsweise des Produktionssystem-Interpreters bestimmt. Der Interpreter realisiert einen „Zwei-Paß-Prozeß“: In einem ersten Schritt wird durch eine zielgerichtete „Aufmerksamkeits-Fokussierung“ der augenblicklich „interessanteste Job“ ausgewählt (wobei AM einen konkreten Begriff von „Interessantheit“ hat); in einem zweiten Schritt werden die für diesen Job relevanten Produktionsregeln zusammengestellt und in der Reihenfolge ihrer Spezifität abgearbeitet, bis die maximal für einen Job zugestandene Rechenzeit (durchschnittlich 30 Sekunden Kernspeicherzeit) verbraucht ist. Sodann wird mit Schritt 1 der „Zwei-Paß-Prozeß“ für einen neuen Job gestartet, bis der Benutzer seine „mathematische Spielsitzung“ beendet. - Zumindest rudimentär sind in diesem KI-System schon so etwas wie „Bewußtseinsfunktionen“ (vgl. Abschnitt 3.3.4) realisiert: Die erwähnte „Aufmerksamkeits-Fokussierung“ wird durch eine entsprechende „Zeigerfunktion“ des Interpreters ermöglicht, während eine „übersetzerfunktion“ das intern erzeugte Verhalten so externalisiert, daß dessen Ergebnisse dem AM-System dialoggesteuert wieder rückgemeldet werden können.

#### 4. Validierung und Anwendbarkeit von Simulationsmodellen

Die Frage der Gültigkeit oder Validität von Simulationsmodellen und deren Anwendbarkeit ist im Grunde nichts anderes als die Frage nach dem Verhältnis von *Theorie* und *Empirie*, wie es sich generell in den Wissenschaften als methodologisches Problem stellt, hier nurmehr konkretisiert auf das Verhältnis der Theorie der Informationsverarbeitung zur Empirie psychischer Phänomene wie Wahrnehmen, Denken, Lernen, Handeln usw. Das heißt dann aber auch, daß es für die Simulationsmethodik keine grundsätzlich anderen, von den übrigen wissenschaftlichen Methoden verschiedenen Probleme der Gültigkeits- und Anwendbarkeitsprüfung gibt. Es gilt lediglich zu bedenken, daß das Paradigma der Computer-Simulation in der Psychologie - wie in Abschnitt 2.1 ausgeführt - „den Typ der dynamischen, deterministischen, qualitativen und analytischen Erkundungsmodelle zur Nachbildung menschlichen Verhaltens am Beispiel von einzelnen und interagierenden Individuen bevorzugt“, wir es also im wesentlichen mit Einzelfalluntersuchungen zu tun haben, so daß viele der gängigen Überprüfungsmethoden - wie z.B. inferenzstatistische Verfahren für statische, stochastische, quantitative Aggregatmodelle - in der Simulationsmethodik wenig anwendbar sind.

Ausgehend von dem in dem Theorie-Empirie-Verhältnis vermittelnden Modellbegriff lassen sich Kriterien entwickeln, nach denen einerseits der Wirklichkeitsbezug, andererseits aber auch der theoretische Status von Simulationsmodellen und KI-Systemen beurteilt werden kann.

## 4.1 Wirklichkeitsbezug und Modellrelationen

Unter erkenntnistheoretisch-methodologischen Aspekten ist das Paradigma der Computer-Simulation in der Psychologie mit den Grundproblemen der psychologischen Meßtheorie vergleichbar: Wirklichkeitsbezug und Modellrelationen bestimmen sich aus dem dreifachen Verhältnis von *Abbildbarkeit*, *Eindeutigkeit* und *Bedeutsamkeit* wissenschaftlicher Beobachtung. Eine Aufschlüsselung dieses Dreiecksverhältnisses wird die Grundlage für eine Diskussion der Validierungs- und Anwendbarkeitsproblematik der Simulationsmethodik abgeben.

### 4.1.1 Modellbildung als *homomorphe Abbildung*

Voraussetzung für die Modellierbarkeit psychischer Phänomene (in der Meßtheorie also die Meßbarkeit psychischer Merkmale wie z.B. „Intelligenz“ oder „Persönlichkeit“) ist (1) die Existenz eines *empirischen Relativs*, formal

$$\text{EmpiR} = \langle M; P_1, \dots, P_k \rangle,$$

in dem bestimmte, empirisch gegebene „Merkmalsträger“  $M$  (z.B. Personen) und „Beziehungen“  $P_i$  zwischen diesen Merkmalsträgern (z.B. Intelligenz oder Persönlichkeit) definierbar sein müssen, und (2) die Konstruierbarkeit eines *theoretischen Relativs*, formal

$$\text{TheoR} = \langle N; R_1, \dots, R_k \rangle,$$

in dem geeignete „theoretische Objekte“  $N$  (z.B. Zahlen) und Relationen  $R_i$  zwischen diesen Objekten (z.B. numerische Prädikate und Operationen) herstellbar sind. Die *Modellbildung* besteht dann aus der homomorphen („strukturhaltenden“) Abbildung  $\varphi$  des empirischen Relativs in das theoretische Relativ, formal

$$\varphi: \text{EmpiR} \rightarrow \text{TheoR},$$

so daß für beliebige  $m, m' \in M$  und  $i = 1, \dots, k$  gilt:

$$\varphi[P_i(m, m')] = R_i[\varphi(m), \varphi(m')] \text{ mit } \varphi(m), \varphi(m') \in N.$$

Mit anderen Worten: Durch den Homomorphismus  $\varphi$  einer Modellbildung bleibt die in den empirischen Relationen  $P_i$  gegebene reale Struktur  $\langle M; P_1,$

$\dots, P_k\rangle$  in der durch die theoretischen Relationen  $R_i$  definierten formalen Struktur  $\langle N; R_1, \dots, R_i \rangle$  erhalten. Am Beispiel des Messens besteht die homomorphe Abbildung aus dem Zuordnen von (ganzen oder reellen) Zahlen zu den (empirischen) Merkmalsträgern in bezug auf die jeweilige Merkmalsausprägung.

Auf das Paradigma der Computer-Simulation angewandt, wird der Homomorphismus einer Modellbildung durch die Programmierung in einer geeigneten Programmiersprache erstellt, wie dies in den Beispielen der Abschnitte 2.2 und 3.2 illustriert wurde. Empirisches Relativ für das in Abschnitt 2.2 eingeführte Beispiel des „Simple Concept Attainment“ war die reale Struktur  $\langle M; P_1, P_2, P_3, P_4 \rangle$  mit

$$\left. \begin{array}{l} M = \{V_l, V_p\} \\ P_1 = \text{Vl-Beispielvorgabe} \\ P_2 = \text{VP-Antwort} \\ P_3 = \text{Vl-Rückmeldung} \\ P_4 = \text{Vl-Kriteriumswert,} \end{array} \right\} \begin{array}{l} \text{vgl.} \\ \text{Flußdiagramm} \\ \text{von Abb. 1} \end{array}$$

das durch die Programmierung in LOGO zugeordnete theoretische Relativ die Programmstruktur  $\langle N; R_1, R_2, R_3, R_4 \rangle$  mit

$$\left. \begin{array}{l} N = \{\text{EXPERIMENTER, SUBJECT}\} \\ R_1 = \text{EXPERIMENTER'S.INSTANCE} \\ R_2 = \text{SUBJECT'S.ANSWER} \\ R_3 = \text{EXPERIMENTER'S.FEEDBACK} \\ R_4 = \text{:CRITERION.VALUE:} \end{array} \right\} \begin{array}{l} \text{vgl.} \\ \text{Programme} \\ \text{in} \\ \text{Abb.2, 4-6 u.a.} \end{array}$$

Für das in Abschnitt 3.2 eingeführte Beispiel des „Single Letter Exclusion“ war empirisches Relativ die reale Struktur  $(M; \{P_i\})$  mit

$$\begin{array}{l} M = \{V_p, \text{Aufgabe des „Unpassenden Streichens“}\} \\ \{P_i\} = \text{die fünf in der Protokollanalyse von Tabelle 2} \\ \quad \text{verwendeten Operatoren,} \end{array}$$

das in Form eines Produktionssystems programmierte theoretische Relativ die Programmstruktur  $\langle N; \{R_i\} \rangle$  mit

$$\begin{array}{l} N = \{\text{SINGLE.LETTER.EXCLUSION, :GIVEN.ITEMS:}\} \\ \{R_i\} = \{\text{SILEX1, . . . , SILEX8}\}. \end{array}$$

#### 4.1.2 Grundprobleme der Modellrelationen

Kennzeichnend für das Verhältnis von Wirklichkeitsbezug und Modellrelationen sind drei Grundprobleme, wie sie in der Theorie des Messens herausgear-

beitet wurden und die sinngemäß auch auf die Theorie der Informationsverarbeitung übertragen werden können:

- (1) Das Abbildbarkeits- oder Repräsentationsproblem.
- (2) Das Eindeutigkeits- oder Transformierbarkeitsproblem.
- (3) Das Bedeutsamkeits- oder Testbarkeitsproblem.

(1) Das *Abbildbarkeitsproblem* beinhaltet die Frage nach der Darstellbarkeit eines Homomorphismus zwischen einem empirischen und einem theoretischen Relativ, wie sie in der obigen Diskussion eingeführt worden ist; Antwort in der Meßtheorie ist die Formulierung eines Abbildbarkeitstheorems (meist aufgrund einer geeigneten Axiomatisierung), in der Theorie der Informationsverarbeitung die Programmierung eines entsprechenden Simulationsmodells. Die oben angeführten Programmbeispiele illustrieren diesen Sachverhalt.

(2) Das *Eindeutigkeitsproblem* stellt die Frage nach der Zulässigkeit von Transformationen eines theoretischen Relativs in ein anderes, um aus den Transformationsbedingungen die Invarianzeigenschaften des gewählten Homomorphismus ablesen zu können. Antwort in der Meßtheorie ist die Angabe der zulässigen numerischen Transformationen, wodurch der Skalentyp des dem Meßvorgang zugrundeliegenden Homomorphismus festgelegt wird. In der Simulationsmethodik würde dem die Neuprogrammierung des ursprünglichen Simulationsmodells - entweder in einer anderen Programmiersprache oder mit einem anderen Modellansatz - entsprechen, ohne daß dadurch das Modellverhalten, der „Trace“ des Simulationsprogramms, verändert wird. Beispiele sind zum einen die ursprüngliche LOGO-Programmierung von SIMPLE.CONCEPT.ATTAINMENT und dessen in Abschnitt 2.3 diskutierte Programmierbarkeit in LISP (vgl. Abb. 17), zum anderen das ursprünglich als Produktionssystem programmierte SINGLE.LETTER.EXCLUSION und dessen in Abschnitt 3.2.3 behandelte sequentielle Programmversion (vgl. Abb. 25); jede der beiden Programmvarianten zeigt ein völlig identisches Modellverhalten, so daß ihre unterschiedliche Programmierung zulässige Transformationen für den jeweils zugrundeliegenden Homomorphismus darstellen.

(3) Das *Bedeutsamkeitsproblem* bezieht sich auf die für die Validitätsprüfung wichtigste Frage der Relevanz theoretischer Aussagen aufgrund einer bestimmten Repräsentation und deren zulässigen Transformationen. Antwort in der Meßtheorie ist die Formulierung konkreter Skalierungsvorschriften und die Angabe zulässiger Rechenoperationen für die Skalenwerte (z.B. zur Berechnung statistischer Kennwerte und zur Anwendung statistischer Tests). Für die Simulationsmethodik wäre eine Antwort in der Verfügbarkeit geeigneter empirischer Tests (wie z.B. der Turing-Test oder der Protokoll-Trace-Vergleich, vgl. Abschnitt 4.3) und in der Ableitbarkeit empirisch prüfbarer Hypo-

thesen aus dem Simulationsmodell (oder aus der Rahmentheorie, die dem Modell zugrunde liegt) zu suchen.

#### 4.1.3 Kommutatives Diagramm

Der Zusammenhang der drei Problembereiche läßt sich in Form eines kommutativen Diagramms darstellen, wie es in Abb. 29 - in Anlehnung an das kommutative Diagramm der Meßtheorie (vgl. Ueckert, 1980c, S. 193) - wiedergegeben ist.

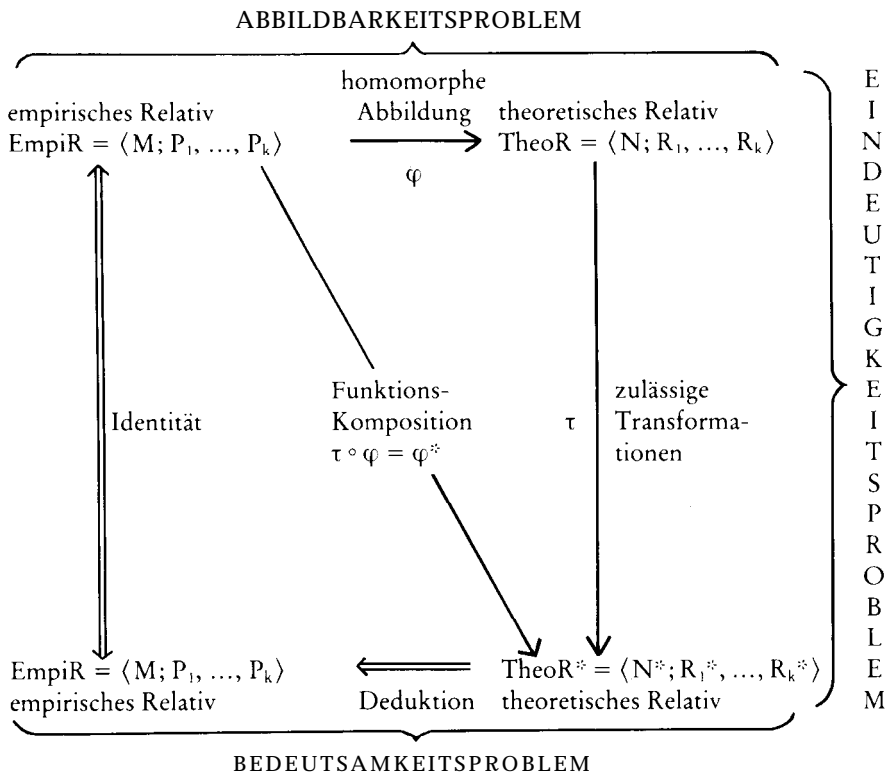


Abb. 29: Kommutatives Diagramm der Modellrelationen für die Computer-Simulation.

Die formal-mathematische Behandlung des Abbildbarkeitsproblems wurde in Abschnitt 4.1.1 schon gegeben. Das Eindeutigkeitsproblem und das Bedeut-

samkeitsproblem beinhaltet den Sachverhalt, daß es zu dem theoretischen Relativ

$$\text{TheoR} = \langle N; R_1, \dots, R_k \rangle$$

stets ein weiteres theoretisches Relativ

$$\text{TheoR}^* = \langle N^*; R_1^*, \dots, R_k^* \rangle$$

mit anderen „theoretischen Objekten“  $N^*$  und Relationen  $R_i^*$  zwischen diesen gibt, das die zulässigen Transformationen in Form einer homomorphen Abbildung

$$\tau: \text{TheoR} \rightarrow \text{TheoR}^*$$

beschreibt, so daß für beliebige  $n, n' \in N$  und  $i = 1, \dots, k$  gilt:

$$\tau[R_i(n, n')] = R_i^*[\tau(n), \tau(n')] \text{ mit } \tau(n), \tau(n') \in N^*.$$

Das aber ist kommutativ äquivalent mit dem Sachverhalt, daß die Funktionskomposition  $\tau \circ \varphi$ , d.h. die Hintereinanderausführung der beiden Abbildungen  $\varphi$  und  $\tau$ , einen neuen Homomorphismus

$$\varphi^*: \text{EmpiR} \rightarrow \text{TheoR}^*$$

erzeugt, der das ursprüngliche empirische Relativ  $\langle M; P_1, \dots, P_k \rangle$  in das neue theoretische Relativ  $\langle N^*; R_1^*, \dots, R_k^* \rangle$  abbildet, so daß gilt:

$$\varphi^* = \{n^* \mid \bigwedge (m \in M) \bigvee (n \in N) \varphi(m) = n \text{ und } \tau(n) = n^*\},$$

d.h. für beliebige  $m \in M$  gilt stets

$$\tau[\varphi(m)] = \varphi^*(m) = n^*.$$

Mit anderen Worten: In einer Modellbildung sind alle die aus ihr deduzierbaren theoretischen Aussagen (Hypothesen) bedeutsam, die die Identität des empirischen Relativs aufgrund der zulässigen Transformationen des theoretischen Relativs erhalten bzw. nicht verletzen, was empirisch in geeigneter Weise - durch entsprechende Tests oder Experimente - nachgewiesen werden kann.

An den bisher eingeführten Programmbeispielen läßt sich dieser Sachverhalt anschaulich illustrieren. Zu dem empirischen Relativ des „Simple Concept Attainment“ wurden in Abschnitt 2.2 und 2.3 zwei verschiedene theoretische Relative diskutiert, einmal die in LOGO programmierte ursprüngliche Version von SIMPLE.CONCEPT.ATTAINMENT (Abb. 2) und zum anderen die entsprechende Version in LISP als SIMPLE-CONCEPT-ATTAINMENT (Abb. 17). Die LISP-Version ist offensichtlich trivial, da sie eine identische Transformation der LOGO-Version darstellt; sie würde zwar ein - für die

Eindeutigkeitsbestimmung gefordertes - gleiches Modellverhalten zeigen, für die Bedeutsamkeitsfrage jedoch wenig hergeben, da sie über das empirische Relativ des „Simple Concept Attainment“ nichts aussagt, was nicht schon in der LOGO-Version - nach einer entsprechenden Gültigkeitsprüfung hierzu - gesagt werden könnte.

Anders verhält es sich dagegen mit dem Beispiel der Aufgabe des „Unpassenden Streichens“ aus Abschnitt 3. Zu dem empirischen Relativ des „Single Letter Exclusion“ wurde das theoretische Relativ des SINGLE.LETTER.EXCLUSION nach der Konzeption von informationellen Produktionssystemen konstruiert, dem in Abb. 25 ein zweites, als sequentielles LOGO-Programm geschriebenes theoretisches Relativ gegenübergestellt wurde; beide Versionen zeigen wiederum ein identisches Modellverhalten und bestimmen damit die zulässigen Transformationen des Simulationsmodells. Unter dem Bedeutsamkeitsaspekt läßt sich hier jedoch ableiten, daß die Modellarchitektur von Produktionssystemen - Arbeitsspeicher, Produktionsspeicher und Interpreter (vgl. Abschnitt 3.2.1) - *nicht* abbildungsrelevant für das Lösen von Aufgaben des „Unpassenden Streichens“ ist (was ja auch nicht das Ziel einer Modellbildung zu dieser Aufgabenstellung ist), denn das sequentielle LOGO-Programm leistet das gleiche wie die Produktionssystemversion, ohne explizit auf Speicherverwaltung und Programminterpretation einzugehen (wenn dies auch implizit dennoch geschieht, vgl. die Diskussion dazu in Abschnitt 3.2.3). Ist die Produktionssystem-Konzeption jedoch selbst - als generelle Modellarchitektur für die menschliche Informationsverarbeitung - Gegenstand der Modellbildung, und zwar im Rahmen einer allgemeinen Theorie der Informationsverarbeitung, dann sind die Fragen der Abbildbarkeit, Eindeutigkeit und Bedeutsamkeit natürlich erneut zu stellen und zu beantworten; eine informelle Behandlung wurde in den Abschnitten 3.2, 3.3 und 3.4 schon gegeben.

## 4.2 Das Eindeutigkeitstheorem von Anderson

Simulationsmodelle in der Psychologie sind nicht nur *statische*, homomorphe Abbildungen psychischer Phänomene, sondern immer auch *dynamische*, prozedurale Nachbildungen der untersuchten Vorgänge, die auf einem Rechner reproduziert werden können. Oder allgemein: Die Theorie der Informationsverarbeitung hat es in ihrer Modellbildung stets mit einem „Repräsentations-Prozeß-Paar“ (Anderson, 1976) zu tun, d.h. jedes in programmierter Form realisierte theoretische Relativ besteht aus einer *Repräsentation* von Information und *Prozessen*, die diese Repräsentation für die Verarbeitung von Information verwenden.

Das Problem ist jedoch, daß die interne Struktur der Repräsentation und die interne Funktionsweise der Prozesse des *empirischen* Relativs, dessen Nachbil-



bildung in dem theoretischen Relativ eines Simulationsmodells angestrebt wird, der direkten Beobachtung nicht zugänglich sind, sondern aus dem beobachtbaren Verhalten erschlossen werden müssen. Diese zunächst das Abbildbarkeitsproblem betreffende Situation charakterisiert Anderson (1976, S. 10-11) folgendermaßen.

Zu jedem Zeitpunkt  $t$  erfolgt eine Eingabe  $i(t)$  in das informationsverarbeitende System, das eine interne Struktur  $s(t)$  mit Hilfe einer *Enkodierfunktion*  $E$  aus dieser Eingabe erzeugt, formal

$$s(t) = E[i(t)].$$

Die Enkodierfunktion

$$E: I \rightarrow S$$

bildet unser Modell der internen Repräsentation von Information in einem informationsverarbeitenden System.

Zum Zeitpunkt  $t' \neq t$  bestimmt die interne Struktur  $s(t')$  mit Hilfe einer *Dekodierfunktion*  $D$  eine Ausgabe  $o(t')$ , die Systemantwort, formal

$$o(t') = D[s(t')].$$

Die Dekodierfunktion

$$D: S \rightarrow O$$

stellt unser Prozeßmodell der Informationsverarbeitung dar.

Ein anschauliches Beispiel ist das Paraphrasieren (Nacherzählen, freie Wiedergabe) von Sätzen:  $i(t)$  wäre der zu paraphrasierende Satz,  $E$  der Vorgang des Verstehens dieses Satzes,  $s(t)$  das damit erzielte (sprachlich-inhaltliche) Verständnis,  $D$  der Prozeß des Erzeugens einer paraphrasierenden Umschreibung des Verstandenen, und  $o(t')$  die (mündliche oder schriftliche) Darstellung der Paraphrase.

Die empirischen Daten, die wir über das beobachtbare Verhalten eines informationsverarbeitenden Systems haben, sind Folgen von  $\langle i(t), o(t') \rangle$ -Paaren (oder in der behavioristischen Terminologie: von Reiz-Reaktions-Paaren) - und das sind in der Tat oft nur „paraphrasierende“ Daten wie beispielsweise die Verbalisierungsprotokolle des „lauten Denkens“ (vgl. Abschnitt 3.1.1). Die Frage ist nun, ob diese Daten hinreichend sind, um die Kodierungsfunktionen  $E$  und  $D$  erschließen zu können. Die Antwort, die Anderson darauf gibt, ist eindeutig „Nein“, und zwar aus folgenden Gründen.

Angenommen, es existiere ein Simulationsmodell  $M$  mit den Kodierungsfunktionen  $E$  und  $D$  (in unserer Terminologie: es gäbe einen Homomorphismus von einem empirischen Relativ in ein theoretisches Relativ). Mit der Definierbarkeit von Äquivalenzklassen  $[i]_E$  unter  $E$ , formal

$$i', i'' \in [i]_E \text{ genau dann, wenn } E(i') = E(i''),$$

lassen sich die zulässigen Transformationen des Simulationsmodells  $M$  in ein anderes Modell  $M^*$  angeben, wenn für die neue Enkodierfunktion  $E^*$ : gezeigt werden kann, daß für alle Eingaben  $i \in I$

$$[i]_{E^*} \subset [i]_E$$

gilt. Da eine Dekodierfunktion  $D^*$ , gegeben die Enkodierfunktion  $E^*$ , immer so gewählt werden kann, daß das resultierende Modellverhalten im Vergleich zwischen  $M$  und  $M^*$  unverändert bleibt, ist es stets möglich, das Originalmodell  $M$  durch ein Zweitmodell  $M^*$  nachzubilden. Anderson beschreibt dies in dem folgenden - hier als *Eindeutigkeitstheorem* bezeichneten - Satz:

Ein Modell  $M$  mit einer Enkodierfunktion  $E$  kann durch ein anderes Modell  $M^*$  mit einer Enkodierfunktion  $E^*$  vollständig nachgebildet werden, wenn für alle Eingaben  $i \in I$  gilt, daß  $[i]_{E^*} \subset [i]_E$ .

Mit anderen Worten: Die Nachbildung von  $M$  ist immer möglich, wenn es dem nachbildenden Modell  $M^*$  in seiner internen Repräsentation gelingt, jede von  $M$  unterscheidbare Eingabe ebenfalls zu unterscheiden. Der triviale Fall wäre, wenn  $M^*$  jeder Eingabe eine eigene - wenn auch u.U. redundante - interne Repräsentation zuordnete; da dies immer erreicht werden kann, gibt es für jedes beliebige  $M$  ein nachbildendes  $M^*$ . (Für einen Beweis des Theorems vgl. Anderson, 1976, S. 11-12.)

Allgemein kann man sagen, daß jedes psychologische Simulationsmodell ein Repräsentations-Prozeß-Paar ist und daß man völlig verschiedene Repräsentationsmodelle (dargestellt durch unterschiedliche Enkodierfunktionen) wählen und dennoch zu äquivalenten Aussagen mit Hilfe der damit konstruierbaren Modelle gelangen kann, da die Repräsentationsunterschiede durch die geeignete Wahl von Prozeßmodellen (dargestellt durch entsprechende Dekodierfunktionen) kompensiert werden können.

### 4.3 Empirische Tests von Simulationsmodellen

Inwieweit die Modellbildung in Form eines Homomorphismus  $\varphi$  bzw.  $\varphi^*$  als gelungen angesehen werden kann (was eine Antwort auf das Abbildbarkeitsproblem der Computer-Simulation darstellen würde) und inwiefern bestimmte Aussagen aus der Realisierung eines Simulationsmodells abgeleitet werden

können (was eine Beantwortung des Bedeutsamkeitsproblems beinhaltet), sind empirische Fragestellungen. Zu ihrer Lösung haben sich in der Simulationsmethodik der Psychologie zwei Überprüfungsverfahren herausgebildet, die insbesondere der Einzelfallcharakteristik psychologischer Simulationsmodelle Rechnung tragen: der sog. Turing-Test des Modellverhaltens und der Protokoll-Trace-Vergleich zwischen Verbalisierungsdaten und Modellausgabe.

#### 4.3.1 Turing-Test

Der Turing-Test ist nach der Intention seines Erfinders, des englischen Mathematikers Turing, eher ein geistreiches Frage-Antwort-Spiel mit dem Rechner als ein ernsthafter Test für Simulationsmodelle (vgl. Turing, 1950). In seiner ursprünglichen Form ist der Turing-Test ein „Imitationsspiel“ derart, daß der Rechner so programmiert ist, einen Menschen in seinem Verbalverhalten perfekt „nachzuahmen“, daß es von dem eines richtigen Menschen nicht mehr zu unterscheiden ist. Konkret sieht der Test - zumindest von der Konzeption her - so aus, daß ein Fragesteller an einem Terminal sitzt und - am besten über zwei getrennte Fernschreiber - sowohl mit dem Rechner als auch mit einem menschlichen Kommunikationspartner verbunden ist; in einem beliebigen strukturierbaren Dialog kann nun der Fragesteller versuchen herauszufinden, welcher seiner beiden Gesprächspartner der Rechner bzw. die Person ist. Spielt man dieses „Imitationsspiel“ mit einer Reihe von Fragestellern durch, dann sollten sich bei der „Maschinenfrage“, d.h. wer der Rechner und wer die Person ist, richtige und falsche Zuordnungen nur nach dem Zufallsprinzip verteilen (d.h. einer Gleichverteilung folgen), wenn das Simulationsprogramm perfekt das nachbildet, was es nachbilden soll.

Es ist klar, daß in dieser unstandardisierten Form der Test wenig brauchbar ist, zumal die ursprüngliche Version von Turing noch etwas komplizierter ist als die oben beschriebene. Danach hat es der Fragesteller entweder mit einer Frau und einem *Mann*, der eine Frau imitiert, oder mit einer Frau und einem *Rechner*, der eine Frau imitiert, zu tun; in beiden Fällen wird am Ende dieses „Imitationsspiels“ die „Frauenfrage“ gestellt: Wer ist jeweils die Frau und wer der eine Frau imitierende Mann bzw. Rechner? In dieser konfundierenden Weise ist der Test natürlich noch weniger brauchbar als in der vereinfachten, auf die „Maschinenfrage“ reduzierten Form. Abelson (1968) hat daher einen „erweiterten Turing-Test“ vorgeschlagen, in dem vor der eigentlichen „Frauenfrage“ die Basisrate für die „Rollenqualität“ des Mannes bestimmt wird, eine Frau in einer bestimmten Dimension (wie z.B. Intelligenz oder Persönlichkeit) zu imitieren; diese Basisrate sollte dann für das Rechnerprogramm ebenfalls erreicht werden (wobei die „Maschinenfrage“ dann ganz entfällt): Liegt das Programm signifikant *über* der Basisrate, ist es „zu männlich“, liegt es *darunter*, ist es „zu weiblich“ in der untersuchten Verhaltensdimension.

Wie alle Rating-Verfahren ist auch dieses wenig zuverlässig, abgesehen von seiner theoretisch recht schwachen Fundierung. Colby und seine Mitarbeiter (vgl. Colby, 1975) entwickelten daher zu ihrem Simulationsmodell des paranoiden Prozesses einen differenzierteren „experimentellen Ununterscheidbarkeitstest“, in dem 10stufige Ratings für die interessierenden Dimensionen (hier: Paranoia) zu transkribierten Rechnerläufen bzw. Verbalprotokollen (hier von paranoiden Patienten, in beiden Fällen als Arzt-Patient-Dialoge) möglich sind. Die Autoren berichten über eine Reihe von statistisch abgesicherten Analysen auf der Grundlage dieses Verfahrens (mit ganz passablen Rater-Übereinstimmungen), so daß dieser „Ununterscheidbarkeitstest“ durchaus als ein praktikabler Ansatz der Validitätsprüfung von Simulationsmodellen angesehen werden kann.

#### 4.3.2 Protokoll-Trace-Vergleich

Grundsätzlich gibt es für jede Modellbildung zwei Arten von Abbildungsfehlern: Das Modell kann zu *wenig* über die abgebildete Realität aussagen, d.h. es bildet nicht all die Aspekte ab, die man in die Modellbildung einbeziehen wollte (*Abbildungsfehler 1. Art*), und das Modell kann zu *viel* über die abgebildete Realität aussagen, d.h. es überzeichnet Aspekte, die in dem abgebildeten Bereich so gar nicht vorkommen (*Abbildungsfehler 2. Art*). Die beste Möglichkeit, solchen Abbildungsfehlern in der Simulationsmethodik auf die Spur zu kommen, ist der Protokoll-Trace-Vergleich zwischen den Verbalisierungsdaten eines Probanden und dem in geeigneter Ausgabeform realisierten Programmlauf des Modells, dem „Trace“ des Rechners (Beispiele für einen solchen „Trace“ sind die Programmläufe in Abb. 3 für das „Simple Concept Attainment“, in Tabelle 3 für das „Single Letter Exclusion“ - hier auch mit entsprechenden Protokolldaten - und in Tabelle 4 für die Produktionssysteme ADD und ADD.TABLE). Im Idealfall sollten Protokoll und „Trace“ in allen die Modellbildung betreffenden Aspekten so übereinstimmen, daß „Vorbild“ und „Nachbild“ nicht mehr unterscheidbar sind (etwa im Sinne des „erweiterten Turing-Tests“ nach Abelson oder des „experimentellen Ununterscheidbarkeitstests“ von Colby).

Im Grunde ist der Protokoll-Trace-Vergleich, wie am Beispiel des „Single Letter Exclusion“ in Abschnitt 3.2 schon gezeigt wurde, konstruktiver Bestandteil der Modellentwicklung. An diesem Beispiel ist auch die besondere Problematik des Verfahrens erkennbar: In jedem Fall handelt es sich um einen *qualitativen* Vergleich zwischen *natürlichsprachlichen* Aussagen und *programmsprachlichen* Ausdrücken, deren inhaltliche Äquivalenz in vielen Fällen durchaus strittig sein kann.

Dennoch eröffnet der Protokoll-Trace-Vergleich einen konkreten Ansatz, Abbildungsfehler in der Entwicklung und Überprüfung von Simulationsmodellen

erkennen und beseitigen zu können. Abweichungen zwischen Protokoll und „Trace“ dahingehend, daß für die Modellbildung relevante Teile von Verbalisierungsdaten im Modell noch nicht reproduziert werden (wie beispielsweise für einen Probanden charakteristische Um- und Irrwege der Informationsverarbeitung), was einem Abbildungsfehler 1. Art entspricht, können für eine entscheidende Modellverbesserung herangezogen werden (weshalb der Protokoll-Trace-Vergleich auch schon in der Phase der Modellentwicklung verwendet wird). Andererseits können Abweichungen zwischen „Trace“ und Protokoll, die sich auf Teile des Programmlaufs beziehen, für die es in den Verbalisierungsdaten keine Entsprechungen gibt (Abbildungsfehler 2. Art), Anlaß für eine Modellanpassung sein, die ein Simulationsmodell „realitätsgerechter“ - und das heißt in den meisten Fällen „weniger künstlich“ oder einfach „menschlicher“ -werden lassen. Dabei ist allerdings zu unterscheiden, ob das Modell lediglich „Verbalisierungslücken“ ausfüllt, die notwendigerweise in einem funktionsfähigen Simulationsmodell überbrückt werden müssen (vgl. das Beispiel des „Single Letter Exclusion“ in Tabelle 3), oder ob es tatsächlich „zu viel“ an Kapazität und Effizienz hinsichtlich der abgebildeten Vorgänge der Informationsverarbeitung bringt, was einem echten Abbildungsfehler 2. Art entspräche.

Reichhaltiges Anschauungsmaterial für die Vorgehensweise des Protokoll-Trace-Vergleichs liefern Newell & Simon (1972) in ihrem Buch über menschliches Problemlösen, in dem eine Fülle von Beispielen (Kryptarithmetik, Logikaufgaben, Schachprobleme) in Protokoll-Trace-Ausschnitten vorgeführt wird. Ein Studium dieser Materialsammlung ist für eine Diskussion der empirischen Testbarkeit von Simulationsmodellen von außerordentlichem Wert.

#### 4.4 Nicht-Falsifizierbarkeit von KI-Systemen

Die gezielte experimentalpsychologische Überprüfung von Simulationsmodellen hat in der Computer-Simulation bisher nur eine untergeordnete Rolle gespielt. Daß hier inzwischen ein Wandel eingetreten ist, beweisen viele neuere, mit der Simulationsmethodik arbeitende Ansätze in der Psychologie (beispielsweise im Bereich der mit semantischen Netzen operierenden Gedächtnistheorien, vgl. die einführende Darstellung von Wender, Colonius & Schulze, 1980).

Besonderer Vorteil der Verwendung von Simulationsmodellen ist die Möglichkeit der unbeschränkten Durchführung von Modellexperimenten, insbesondere mit verschiedenen Modellvarianten: Jede dieser Modellvarianten kann mit den empirischen Daten experimentalpsychologischer Untersuchungen verglichen werden, bis am Ende eine Variante resultiert, die als optimales Modell des nachgebildeten Realitätsbereichs angesehen werden kann. Ein Beispiel ist das

in Abschnitt 2.2 mit vier Modellvarianten vorgestellte Programm des „Simple Concept Attainment“. In ihrer diesbezüglichen Arbeit haben Gregg & Simon (1967) eine ausführliche Diskussion der Modellvarianten im Vergleich mit empirischen Daten aus einer Reihe von Experimenten geliefert, deren Ergebnis nicht zuletzt darin zu sehen ist, daß die Simulationsmethodik sowohl der verbalsprachlichen als auch der mathematischen Modellbildung in Präzision (qua modelltheoretischer Annahmen), Expliztheit (qua Programmerstellung) und Validität (qua Experiment-Modell-Vergleich) überlegen ist.

Dennoch, trotz aller empirischer Testbarkeit mit Turing-Test oder Protokoll-Trace-Vergleich und experimentalpsychologischer Überprüfbarkeit durch Experiment-Modell-Vergleich kann man sich die Frage stellen, ab wann ein Simulationsmodell oder gar die dahinterstehende Theorie der Informationsverarbeitung als verifiziert - oder nach wissenschaftstheoretischen Überlegungen richtiger als falsifiziert - angesehen werden kann. Aus der Logik von Maschinen, die per Konstruktionsprinzip zur Informationsverarbeitung in der Lage *sind*, läßt diese Frage nur eine Antwort zu: Jedes lauffähige Computer-Programm ist - gleichgültig, ob es einen realen Prozeß der Informationsverarbeitung nachbildet oder nicht - eine Anwendung, eine konkrete Realisation der *Theorie* der Informationsverarbeitung und kann von daher grundsätzlich nicht als Falsifikationsinstanz für eben diese Theorie angesehen werden. Mit anderen Worten: Systeme der „künstlichen Intelligenz“ - im weiten wie im engeren Sinne - sind prinzipiell *nicht falsifizierbar*, sondern allenfalls in unterschiedlichem Umfang auf die Modellierung realer Prozesse *anwendbar*.

Die einen derartigen Anspruch rechtfertigende Theoriekonzeption, der sog. strukturalistische Theoriebegriff oder „non-Statement view“ von Theorien (vgl. Stegmüller, 1973), soll im folgenden auf die mit der Computer-Simulation verbundene Theorie der Informationsverarbeitung, insbesondere im Zusammenhang mit der Modellarchitektur von informationellen Produktionssystemen, direkt bezogen und diskutiert werden.

#### 4.4.1 Der strukturalistische Theoriebegriff

Nach der herkömmlichen wissenschaftstheoretischen Auffassung besteht eine Theorie aus einem System von Sätzen (oder „Aussagen über die Realität“), deren Zusammenhang untereinander durch logische Ableitungsbeziehungen (Widerspruchsfreiheit, Kohärenz u. a.) und deren Gültigkeitsanspruch durch Wahrheitskriterien (Bestätigung, Falsifikation, Korrespondenz mit der Realität usw.) geregelt ist. Dieser als „Aussagenkonzeption“ bezeichneten Auffassung von Theorien stellte Sneed einen anderen, ursprünglich im Bereich der theoretischen Physik entwickelten Ansatz gegenüber, den sog. „non-statement view“ von Theorien, den Stegmüller (1973, 1980) auf die aktuelle wissen-

schaftstheoretische Diskussion übertragen hat (insbesondere vor dem Hintergrund der Kuhnschen Thesen über „normalwissenschaftlichen Fortschritt“ und „revolutionären Wandel der Wissenschaft“). Nach diesem - im folgenden als „strukturalistische Theoriekonzeption“ bezeichneten - Ansatz bestehen Theorien nicht aus Satzsystemen mit logischen Ableitungsbeziehungen und wahrheitsdefiniten Gültigkeitskriterien, sondern aus logisch-mathematischen Konstruktionen, deren instrumenteller Gebrauch durch modelltheoretische Formulierungen geregelt wird.

Ohne auf Einzelheiten allzusehr einzugehen (man vergleiche dazu die ausführliche Darstellung von Stegmüller, 1973), wird im folgenden die Übertragbarkeit dieses Ansatzes auf die Theorie der Informationsverarbeitung versucht, um damit eine Argumentationsbasis zu gewinnen, auf der die Anwendbarkeit und der instrumentelle Gebrauch der Computer-Simulation und der „künstlichen Intelligenz“-Forschung in plausibler Weise gerechtfertigt werden kann.

Nach der strukturalistischen Theoriekonzeption besteht eine Theorie T untrennbar aus zwei Komponenten:

- (1) einer *logischen* Komponente K, die den „Kern“ der diese Theorie kennzeichnenden Struktur in logisch-mathematischen Kategorien beschreibt, und
- (2) einer *empirischen* Komponente A, die die „intendierten Anwendungen“ der Theorie auf konkrete Gegenstandsbereiche der Realität beinhaltet.

Formal ist jede Theorie durch das geordnete Paar

$$T = (K, A)$$

darstellbar, so daß es grundsätzlich nicht zulässig ist, in einer Theorie nur von der einen Komponente zu reden, ohne die andere zu erwähnen, und umgekehrt. Oder allgemein gesagt: Es gibt keine Theorie ohne Anwendungen und es gibt keine Anwendungen ohne entsprechende Theorie.

In erster Annäherung kann man die in Abschnitt 4.1 eingeführten Begriffe des empirischen und des theoretischen Relativs auf diesen Theoriebegriff beziehen: Theoretische Relative gehören zum Strukturkern K, empirische Relative zu den Anwendungen A einer bestimmten Theorie T.

Kennzeichnend für den strukturalistischen Theoriebegriff ist der Sachverhalt, daß eine Theorie einerseits bei *gleichem* Strukturkern *verschiedene* intendierte Anwendungen haben kann (beispielsweise wenn die Individuenbereiche dieser Anwendungen verschieden sind), die durch *allgemeine*, in allen diesen Anwendungen geltenden Gesetzmäßigkeiten oder „Nebenbedingungen“ miteinander verbunden sind, und andererseits der Strukturkern *so erweitert* werden kann, daß spezielle Gesetze nur in bestimmten, nicht jedoch in anderen Anwendun-

gen gelten, was durch die Angabe von *speziellen* Nebenbedingungen geregelt wird. Auf die Theorie der Informationsverarbeitung übertragen heißt das, daß mit der Formulierung eines informationsverarbeitenden Systems (IVS) zunächst der Strukturkern  $K(\text{IVS})$  für die Theorie festgelegt wird, als deren intendierte Anwendungen  $A(\text{IVS})$  primär der Rechner, sekundär aber auch - durch die Entwicklung der Computer-Simulation bedingt - der Mensch angesehen wird (also zwei verschiedene, mit unterschiedlichen Individuenbereichen arbeitende Anwendungen des gleichen Strukturkerns). Darüber hinaus gibt es jedoch auch spezielle Anwendungen der Theorie der Informationsverarbeitung, die im Falle des Menschen zu Eigengesetzlichkeiten führt, die von denen eines Rechners verschieden sind.

#### 4.4.2 Die logische Komponente der Theorie der Informationsverarbeitung

Jedes Modell der Theorie der Informationsverarbeitung, d.h. jedes konkrete informationsverarbeitende System IVS, setzt sich aus folgenden Systemteilen zusammen:

- (1) Eingabe-/Ausgabe-Einheiten,
- (2) Datenspeicher,
- (3) Programmspeicher,
- (4) Prozessor für Daten-/Programmspeicher.

In der Modellarchitektur von informationellen Produktionssystemen (vgl. Abschnitt 3.2.1) entsprechen den Systemteilen (2) bis (4) Arbeitsspeicher, Produktionsspeicher und Interpretier.

Über diesen vier Systemteilen läßt sich die logische Komponente der Theorie der Informationsverarbeitung, der Strukturkern  $K(\text{IVS})$ , systematisch aufbauen.

Auf den Eingabe-/Ausgabe-Einheiten operieren zwei Mengen

$I$  := nicht-leere („empirische“) Menge von *Eingaben* („Inputs“) für ein IVS (einschließlich der leeren Eingabe),

$O$  := nicht-leere („empirische“) Menge von *Ausgaben* („Outputs“) für ein IVS (einschließlich der leeren Ausgabe),

deren Verknüpfung durch eine Funktion geregelt wird:

$R: I \rightarrow O$  := nicht-theoretische („empirische“) Funktion („*Response-Funktion*“) des IVS, die die Menge der Eingaben in die Menge der Ausgaben abbildet, so daß für beliebige benachbarte Zeitpunkte  $t < t'$  gilt:

$$R[i(t)] = o(t').$$



Die Eingabe-/Ausgabemengen bezeichnen die für das IVS empirisch verfügbare Information („Reiz-Reaktions-Muster“ der behavioristischen Psychologie). Die Response-Funktion  $R$  ist eine dynamische Funktion dahingehend, daß die Systemausgabe  $o(t')$  stets mit einer gewissen Latenzzeit  $t'-t$  auf die Systemeingabe  $i(t)$  folgt. Diese Funktion ist insofern „nicht-theoretisch“, als sie sich empirisch stets aus der Auswertung von  $\langle i(t), o(t') \rangle$ -Paaren, also aus dem Eingabe-Ausgabe-Verhalten eines IVS, bestimmen läßt. Die Einbeziehung der leeren Eingabe bzw. Ausgabe ist notwendig, um in der Zeitvariable  $t$  keine „Lücken“ offen zu lassen. (D.h. es wird nicht ausgeschlossen, daß auf eine Eingabe unmittelbar keine Ausgabe folgt oder für eine Ausgabe unmittelbar keine Eingabe gegeben ist oder aber daß das IVS zeitweise auch nach außen hin inaktiv ist, also ein leeres Eingabe-Ausgabe-Verhalten zeigt.)

Mit diesen Definitionen ist eine Menge von Systemen gegeben, die ausschließlich empirisch bestimmt sind und die als Menge der „partiellen potentiellen Modelle“

$$M_{pp} = \{x \mid x = \langle R; I, O \rangle\}$$

die erste Komponente des Strukturkerns  $K(\text{IVS})$  darstellen; „partiell“ heißen diese Modelle deshalb, weil sie noch keine theoretischen Systemgrößen enthalten, und „potentiell“, weil sie nur „mögliche“, nicht jedoch auch schon „wirkliche“ Modelle eines IVS sind. Die Modelle  $x \in M_{pp}$  sind dynamische Modelle, da das Modellverhalten durch die empirische Response-Funktion zeitabhängig ist.

Auf dem Datenspeicher - oder dem Arbeitsspeicher von informationellen Produktionssystemen - sind weitere Entitäten definierbar, die in bezug auf die Theorie der Informationsverarbeitung theoretischen Status haben, weil sie der unmittelbaren Beobachtung nicht zugänglich sind. Im einzelnen sind dies die Menge

$S :=$  nicht-leere („theoretische“) Menge von *internen Repräsentationen* („Symbolstrukturen“) eines IVS zur systemeigenen Darstellung von Information,

und die Systemfunktionen

$E: I \rightarrow S :=$  theoretische („nicht-empirische“) Funktion („*Enkodierfunktion*“) des IVS, die die Menge der Eingaben in die Menge interner Repräsentationen abbildet, so daß für jeden beliebigen Zeitpunkt  $t$  gilt:

$$E[i(t)] = s(t);$$

$D: S \rightarrow O :=$  theoretische („nicht-empirische“) Funktion („*Dekodierfunktion*“) des IVS, die die Menge der internen Repräsentationen in die Menge der Ausgaben abbildet, so daß für jeden beliebigen Zeitpunkt  $t$  gilt:

$$D[s(t)] = o(t);$$

$U: S \rightarrow S :=$  theoretische („nicht-empirische“) Funktion („*Umstrukturierungsfunktion*“) des IVS, die die Menge der internen Repräsentationen in sich selbst abbildet, so daß für beliebige benachbarte Zeitpunkte  $t < t'$  gilt:

$$U[s(t)] = s'(t').$$

Die Repräsentationsmenge  $S$  und die Kodierungsfunktionen  $E$  und  $D$  sind aus der Darstellung des Eindeutigkeitstheorems von Anderson (vgl. Abschnitt 4.2) schon bekannt und entsprechen der dort eingeführten Bedeutung. Die beiden Kodierungsfunktionen sind statische Funktionen, da sie ihre Werte jeweils zeitgleich aus den Argumenten erzeugen. Die Umstrukturierungsfunktion  $U$  dagegen ist eine dynamische Funktion, da jede Umstrukturierung mit einem gewissen Zeitaufwand  $t'-t$  verbunden ist. Diese Funktion dient sowohl der Darstellbarkeit von sequentiellen Vorgängen der Informationsverarbeitung mittels

$$U[s(t)] = s'(t')$$

als auch der Erklärbarkeit von Eingaben, aus denen unmittelbar keine Ausgabe erfolgt, mittels

$$U[E[i(t)]] = s(t'),$$

und von Ausgaben, für die es keine direkte Eingabe gibt, mittels

$$D[U[s(t)]] = o(t').$$

Auf dem Programmspeicher - oder dem Produktionsspeicher von Produktionssystemen - sind damit drei Mengen von „*Programmstrukturen*“ definierbar:

$E(I) :=$  Menge der Verknüpfungen der Enkodierfunktion mit ihrer Eingabemenge;

$D(S) :=$  Menge der Verknüpfungen der Dekodierfunktion mit der zu Systemausgaben geeigneten Repräsentationsmenge;

$U(S) :=$  Menge der Verknüpfungen der Umstrukturierungsfunktion mit der zu internen Umstrukturierungen verfügbaren Repräsentationsmenge.

Die Programme eines IVS bestehen ausschließlich aus der geeigneten Zusammenstellung von Programmstrukturen dieser drei Verknüpfungsmengen.

Schließlich operiert auf dem Prozessor für Daten- und Programmspeicher - dem Interpreter von Produktionssystemen - eine Meta-Funktion

$P: \{„E(I)“, „D(S)“, „U(S)“\} \rightarrow \{E(I), D(S), U(S)\} :=$  theoretische („nicht-empirische“) Funktion höherer Ordnung („*Prozessorfunktion*“) des IVS, die die Programmausdrücke der Form „ $F(X)$ “ derart in Programmausführungen  $F(X)$  übersetzt, daß für beliebige  $F(X) = E(I)/D(S)/U(S)$  gilt:

$$P[„F(X)“] = F(X).$$

Die Prozessorfunktion  $P$  ist insofern eine Metafunktion, als sie die konkrete Interpretation der drei Systemfunktionen  $E$ ,  $D$ ,  $U$  auf ihren jeweiligen Argumentmengen nach Maßgabe der im Programmspeicher gegebenen Programmstrukturen regelt.

Mit diesen zusätzlichen Definitionen ist nunmehr eine spezifischere Menge von Systemen angebar: Neben empirischen Systemgrößen enthalten sie auch solche, die für die interne Struktur eines IVS und deren Interpretation kennzeichnend sind. Diese Systeme bilden als die Menge der „*potentiellen Modelle*“

$$M_p = \{y \mid y = \langle R, E, D, U, P; I, O, S \rangle\}$$

die *zweite* Komponente des Strukturkerns  $K(\text{IVS})$ ; auch sie sind nur „mögliche“ und noch nicht „wirkliche“ Modelle eines IVS, da der Zusammenhang zwischen empirischen und theoretischen Systemgrößen noch nicht vollständig spezifiziert ist. Auch die Modelle  $y \in M_p$  sind dynamische Modelle, da sie neben der empirischen Zeitfunktion  $R$  auch noch die theoretische Zeitfunktion  $U$  enthalten.

Schließlich wird über eine Menge von Systemen der Zusammenhang zwischen der empirischen Response-Funktion  $R$  und den theoretischen Systemfunktionen  $E$ ,  $D$ ,  $U$  so konkretisiert, daß jede Ausgabe des IVS eindeutig aus den Kodierungs- und/oder Umstrukturierungsoperationen von Eingaben bestimmbar ist; es ist dies die Menge der *eigentlichen Modelle*

$$M = \{z \mid z \in M_p \text{ und } R[i(t)] = D[U[E[i(t)]]]\}.$$

Diese Modelle bilden die *dritte* Komponente des Strukturkerns  $K(\text{IVS})$ ; sie sind die „wirklichen“ Modelle eines IVS dahingehend, daß in ihnen alle Systemgrößen vollständig spezifiziert und in Zusammenhang gebracht sind, die für ein eindeutiges Systemverhalten relevant sind. Die Modelle  $z \in M$  erklären die über die empirische Response-Funktion  $R$  beobachtbaren Latenzzeiten  $t'-t$  aus der Zeitverschiebung von Umstrukturierungsoperationen durch die theoretische Zeitfunktion  $U$ ; damit sind auch diese Modelle als dynamische Modelle anzusehen.

Als letztes sind noch zwei weitere Komponenten des Strukturkerns der Theorie der Informationsverarbeitung zu definieren:

(1) Eine „*Restriktionsfunktion*“

$$q: M_p \rightarrow M_{pp},$$

die die Menge der potentiellen Modelle in die Menge der partiellen potentiellen Modelle derart abbildet, daß gilt:

$$q(\langle R, E, D, U, P; I, O, S \rangle) = \langle R; I, O \rangle,$$

d.h. die Restriktionsfunktion ermöglicht den Nachweis, ob ein Modell der Theorie der Informationsverarbeitung tatsächlich theoriespezifische Funktionen enthält (d.h. ob  $M_p \neq M_{pp}$  gilt) oder nicht (d.h. ob  $M_p = M_{pp}$ , so daß die als „theoretisch“ angesehenen Systemfunktionen nichts anderes als identische Transformationen sind).

(2) Eine Angabe von *allgemeinen Nebenbedingungen*, die den theoretischen Funktionen (und nur diesen) der potentiellen Modelle  $M_p$  auferlegt werden; für die Theorie der Informationsverarbeitung ist dies die äquivalenzklassenbildende Nebenbedingung

$$N(M_p) = \langle \approx, = \rangle$$

für die Funktionen E, D, U und P derart, daß gilt:

$$i', i'' \in [i]_E \Leftrightarrow E(i') = E(i'') \text{ (Enkodierungs-Äquivalenz)}$$

$$s', s'' \in [s]_D \Leftrightarrow D(s') = D(s'') \text{ (Dekodierungs-Äquivalenz)}$$

$$s', s'' \in [s]_U \Leftrightarrow U(s') = U(s'') \text{ (Umstrukturierungs-Äquivalenz)}$$

$$„F'(A)“, „F''(B)“ \in „F(X)“_P \Leftrightarrow F'(A) = F''(B)$$

(Interpretations-Äquivalenz)

Mit anderen Worten: Jedes IVS muß in der Lage sein, äquivalenten Eingaben eine identische interne Repräsentation zu geben bzw. aufgrund äquivalenter Repräsentationsstrukturen zu identischen Ausgaben und/oder identischen Umstrukturierungen zu kommen. Und umgekehrt: Ein IVS kann nur aufgrund identischer interner Repräsentationen Eingaben als äquivalent und nur aufgrund identischer Ausgaben bzw. Umstrukturierungen Repräsentationsstrukturen als äquivalent erkennen. Ferner müssen äquivalente Programmausdrücke zu einer identischen Interpretation durch das IVS führen (und umgekehrt).

Mit diesen fünf Komponenten ist nunmehr der Strukturkern der Theorie der Informationsverarbeitung formal definiert als

$$K(IVS) = \langle M_{pp}, M_p, M; Q, N(M_p) \rangle.$$

Die bisher diskutierten Systemgrößen eines informationsverarbeitenden Systems lassen sich anhand der Programmiersprache LOGO und den in ihr eingeführten Programmbeispielen leicht veranschaulichen.

Eingabe-Einheit ist die Tastatur des Fernschreibers oder des Sichtgerätes, Ausgabe-Einheit Papier oder Bildschirm. Eingaben erfolgen in laufenden LOGO-Programmen mit der LOGO-Operation ‚request‘, Ausgaben mit der LOGO-Anweisung ‚print‘ (oder anderer, in LOGO verfügbarer Ausgabefunktionen). Der empirischen Systemfunktion  $R[i(t)]$  entspricht also im einfachsten Falle die LOGO-Programmzeile ‚print request‘ (was die identische Ausgabe einer

Eingabe bedeutet). Für das in Abschnitt 3.4 eingeführte ADD-Produktionssystem entspräche der empirischen Systemfunktion beispielsweise die Folge

,RUN „ADD“ \*3, \*2‘ mit der Ausgabe ,5‘.

In LOGO dient ein Teil des Arbeitsspeichers, der sog. Variablenspeicher, als Datenspeicher, während der Rest des Arbeitsspeichers den Programmspeicher darstellt (wobei Daten und Programme auch langfristig in sog. Dateien gespeichert werden können). Interne Repräsentationsstrukturen sind in LOGO als variable Namen+Wert-Zuweisungen (in Form von LOGO-Wörtern und -Sätzen) darstellbar. Beispiele für die Grundform theoretischer Systemfunktionen in LOGO sind:

Enkodierung  $E[i(t)]$ : make „X“ request

Dekodierung  $D[s(t)]$ : print :X:

Umstrukturierung  $U[s(t)]$ : make „Y“ :X:

Operationssequenzen anhand dieser Beispiele wären dann:

$U[E[i(t)]]$ : make „X“ request , make „Y“ :X:

$D[U[s(t)]]$ : make „Y“ :X: , print :Y:

$D[U[E[i(t)]]]$ : make „X“ request , make „Y“ :X: , print :Y:

Die weitaus größte Zahl von Operationen und Anweisungen in einem LOGO-Programm sind in der Regel Umstrukturierungsoperationen an internen Repräsentationsstrukturen, in „lernfähigen“ Programmen aber auch an Programmstrukturen selbst, wie die EXPAND-Funktion für das ADD.TABLE-Produktionssystem in Abschnitt 3.4 zeigt.

Die Bedeutung des Prozessors für Daten- und Programmspeicher läßt sich am Beispiel des in Abschnitt 3.2 eingeführten Interpreters von informationellen Produktionssystemen gut demonstrieren. Im Grunde ist die ,do‘-Anweisung die Prozessorfunktion von LOGO (der top-level von LOGO ist nichts anderes als eine endlose ,do request‘-Schleife), und wie aus der Programmierung des Interpreters in Abb. 23 und 24 zu ersehen ist, spielt hier-vor allem in der READY- und in der FIRE-Funktion von PROCESS - die ,do‘-Anweisung die zentrale Rolle für die Interpretation von Produktionssystemen und von Produktionsregeln. (Dabei ist es sogar möglich, daß die Prozessorfunktion ,do‘ selbst im Aktionsteil von Produktionsregeln wieder vorkommen kann, wie die Produktionen SILEX4 von SINGLE.LETTER.EXCLUSION und ADD1, ADD5 und TAB1 von ADD bzw. ADD.TABLE zeigen.) Anhand der oben eingeführten LOGO-Beispiele für die Systemfunktionen E, D, U wäre für die Prozessorfunktion P zu schreiben:

$P[,E[i(t)]“]$ : do „make „X“ request“

$P[,D[s(t)]“]$ : do „print :X:“

$P[,U[s(t)]“]$ : do „make „Y“ :X:“

Und für die Operationssequenzen:

$P[„U[E[i(t)]]“]:$  do „make „X“ request , make „Y“ :X:“

$P[„D[U[s(t)]]“]:$  do „make „Y“ :X: , print :Y:“

$P[„D[U[E[i(t)]]]“]:$  do „make „X“ request , make „Y“ :X: , print :Y:“

wobei die Anführungszeichen zu Beginn und am Ende einer ‚do‘-Anweisung von solchen innerhalb der ‚do‘-Anweisung selbstverständlich zu unterscheiden sind (und von LOGO tatsächlich auch unterschieden werden).

#### 4.4.3 Die empirische Komponente der Theorie der Informationsverarbeitung

Die empirische Komponente A(IVS) der Theorie der Informationsverarbeitung beinhaltet die Menge der „intendierten Anwendungen“ dieser Theorie auf reale informationsverarbeitende Systeme, zu denen vor allem - wie bereits erwähnt - einerseits der Rechner, andererseits aber auch der Mensch zählt.

Unter Bezugnahme auf die im vorigen Abschnitt vorgestellte logische Komponente K(IVS) besteht die Menge A(IVS) aus einer konkret spezifizierbaren, *empirisch benennbaren* Teilmenge der partiellen potentiellen Modelle  $M_{pp}$ , formal

$$A(IVS) \subset M_{pp},$$

d.h. derjenigen Komponente des Strukturkerns K(IVS), die in ihren Individuenbereichen (der Eingabemenge 1 und der Ausgabemenge 0) und deren funktionaler Verknüpfung (mittels der Response-Funktion R) ausschließlich empirisch bestimmt ist. Die spezifizierende Bedingung der „empirischen Benennbarkeit“ von intendierten Anwendungen A(IVS) besagt, daß konkrete informationsverarbeitende Systeme angebbar sein müssen, ohne in deren Beschreibung auf die Theorie der Informationsverarbeitung selbst einzugehen. Im Falle des Rechners heißt das, seine „hardware“ zu beschreiben (zentrale Recheneinheit, periphere Speicher, Eingabe-Ausgabe-Geräte usw.) und die Implementierbarkeit bestimmter Programmiersprachen anzugeben (Alphabet, Grundvokabular, Syntax und Grammatik). Für den Menschen heißt das, ihn in seiner Gegenständlichkeit als psychologisches Subjekt zu beschreiben: als Lebewesen in einem biologischen und sozialen Lebenszusammenhang mit beobachtbaren Reaktionsformen („Verhalten“), Aktionsmöglichkeiten („Handeln“) und Erlebnisweisen („Motivationen und Kognitionen“).

Im Gegensatz zu der expliziten Bestimmbarkeit (durch Aufzählung) oder der impliziten Definierbarkeit (durch Angabe notwendiger und hinreichender Merkmale) von partiellen potentiellen Modellen (vgl. die definitorische Einführung von  $M_{pp}$  in Abschnitt 4.4.2) sind die intendierten Anwendungen A(IVS) in der Regel nur als „paradigmatische Beispiele“ B, als „typische

Exemplare“ informationsverarbeitender Systeme angebbbar (vgl. Stegmüller, 1973, S. 195-207). Insbesondere muß es - für den Erfinder einer Theorie T bzw. für all diejenigen, die diese Theorie akzeptieren - eine *paradigmatische Beispielmenge*  $B_o \subset A$  (mit  $B_o \subset B$ ) geben, die *unverzichtbarer* Bestandteil für die Anwendbarkeit der Theorie T ist, so daß letztlich der nicht-reduzierbare Umfang einer Theorie mit

$$T = \langle K, B_o \rangle$$

formulierbar ist.

Die Frage ist nun: Was ist die paradigmatische Beispielmenge  $B_o(IVS)$  für die Theorie der Informationsverarbeitung? Eine umfassende Antwort auf diese Frage läßt sich sicher nicht geben, doch dürften sich - für den Bereich der Computer-Simulation und der KI-Forschung - jene frühen Modelle dazu zählen lassen, die beispielsweise bereits in dem wegweisenden Buch von Feigenbaum & Feldman (1963) versammelt sind: Maschinen, die

Schach (Newell, Shaw & Simon) und Dame (Samuel) spielen,  
Theoreme aus Logik (Newell, Shaw & Simon) und Geometrie (Gelernter u. Mitarb.) beweisen,  
natürlichsprachliche Fragen beantworten (Green u. Mitarb., Lindsay),  
visuelle Muster erkennen (Selfridge & Neisser, Uhr & Vossler),  
Probleme lösen (Newell & Simon),  
sinnlose Silben (Feigenbaum) und sinnvolle Begriffe (Hunt & Hovland) lernen,  
Entscheidungen unter Unsicherheit treffen (Feldman, Clarkson),  
ja, sogar interpersonelles soziales Verhalten nachbilden (Gullahorn & Gullahorn).

Für den Bereich menschlicher Systeme der Informationsverarbeitung läßt sich die paradigmatische Beispielmenge  $B_o(IVS)$  weniger eindeutig angeben. Mit Sicherheit dazu zählen kann man all jene empirischen Versuche, die schon am Anfang der Computer-Simulation die Grundlage für die Modellentwicklung (wie beispielsweise für den „Logic Theorist“ und den „General Problem Solver“ von Newell, Shaw & Simon, 1957, 1959) bildeten. Man kann aber, auch wenn der Begriff der Information bzw. eine Theorie der Informationsverarbeitung noch unbekannt waren, all die früheren experimentalpsychologischen Untersuchungen, die im Gefolge des Behaviorismus entstanden (bis hin, wenn man will, zu den ersten psychophysischen Versuchen im 19. Jahrhundert), als paradigmatische Beispiele ansehen, sofern nur für das beobachtbare Eingabe-Ausgabe-Verhalten immer eine empirische Response-Funktion bestimmbar bleibt. Gerechterweise sollte man aber den Menschen nur unter jenen Aspekten als Paradigma der Informationsverarbeitung verstehen, die explizit auf eine entwickelte Theorie in dem hier dargestellten Sinne Bezug nehmen.

#### 4.4.4 Der instrumentelle Gebrauch der Theorie der Informationsverarbeitung

Nach der strukturalistischen Theoriekonzeption konkretisiert sich der Umgang mit einer Theorie T auf die Frage, wie aufgrund eines gegebenen Strukturkerns K aus den intendierten Anwendungen A genau jene spezifiziert werden können, die für die augenblickliche Betrachtung interessieren. In der Regel gelten in bestimmten Anwendungen ganz spezifische Gesetzmäßigkeiten, die für andere Anwendungen nicht zutreffen, so daß deren Besonderheit durch die Angabe spezieller Nebenbedingungen herausgearbeitet werden muß. Dabei dient der vorgegebene Strukturkern, in dem die *allgemeinen*, in allen Anwendungen geltenden Nebenbedingungen formuliert sind, als Instrument für die Herausarbeitung der *speziellen*, nur in einer bestimmten Anwendung gültigen Nebenbedingungen.

Für den Bereich der Computer-Simulation (und auch der KI-Forschung) gilt beispielsweise als spezielle Nebenbedingung, daß sowohl die Eingabemenge I als auch die Ausgabemenge O eines IVS *Teilmenge* der internen Repräsentationsmenge S ist, deren Alphabet und Grundvokabular zudem noch von der verwendeten Programmiersprache abhängig ist. Für die Gültigkeit von Simulationsmodellen heißt das dann, daß diese sich notwendigerweise auf die Nachbildung menschlichen Sprachverhaltens bzw. auf die Darstellung sprachlich beschreibbaren nonverbalen Verhaltens beschränken müssen.

Für den Menschen als Gegenstand der Theorie der Informationsverarbeitung stehen dem andere spezielle Nebenbedingungen entgegen, die die Besonderheit der menschlichen Informationsverarbeitung bestimmen. Vor allem die Mehrkanaligkeit (Parallelität) der Enkodierung und Dekodierung von Information, möglicherweise aber auch eine Multiplizität der Repräsentation von Information im Gehirn bedingen Eigengesetzlichkeiten, die sich in einer Vielfalt psychischer Phänomene niederschlagen und sich von daher von den Eigengesetzlichkeiten eines Rechners unterscheiden werden. Zweifellos gibt es auch eine Reihe von Bewußtseinsfunktionen, die im Bereich der menschlichen Informationsverarbeitung über die einfache Prozessorfunktion der kognitiven Exekutive eines IVS hinausgehen (vgl. dazu Ueckert, 1980b).

Die konkrete Herausarbeitung solcher spezieller Nebenbedingungen - methodisch in Form von geeigneten Erweiterungen des Strukturkerns der Theorie (vgl. Stegmüller, 1973, S. 122-139) - ist Aufgabe einzelwissenschaftlicher Untersuchungen. Die Theorie der Informationsverarbeitung ist nach der strukturalistischen Theoriekonzeption nur die *formale* Rahmentheorie (vergleichbar etwa der Meßtheorie für die „Meßbarkeiten des Psychischen“), deren instrumenteller Gebrauch die Entwicklung inhaltlich-psychologischer Theorien für die einzelnen Gegenstandsbereiche menschlicher Informationsverarbeitung erleichtert.



Das Problem der Falsifizierbarkeit von Simulationsmodellen reduziert sich hierbei auf die Frage, ob ein mit der Modellentwicklung unternommener Anwendungsversuch der Theorie der Informationsverarbeitung als *erfolgreich* anzusehen ist oder nicht, d.h. ob ein konkret vorgegebenes Simulationsmodell noch zu der Menge der intendierten Anwendungen A(IVS) gehört. Im negativen Fall ist nicht die Theorie der Informationsverarbeitung - weder für den Menschen noch für den Rechner - falsifiziert, sondern nur der Versuch ihrer Anwendung an dem Modell gescheitert, was aber nicht ausschließt, daß ein erneuter, mit einem verbesserten Modell arbeitender Anwendungsversuch nicht doch noch erfolgreich sein wird.

In ihrer Übertragung auf den Rechner ist die Theorie der Informationsverarbeitung bisher stets erfolgreich angewendet worden, sofern die Programme das intendierte Modellverhalten zeigen konnten. In diesem Zusammenhang ist die Theorie der Informationsverarbeitung sogar prinzipiell nicht falsifizierbar: Die theoretischen Systemfunktionen eines potentiellen Modells der Informationsverarbeitung können in einem Rechner jederzeit in dem Sinne „empirisch“ gemacht werden, daß ihre Implementierung und Arbeitsweise durch geeignete Modellausgaben (z. B. über entsprechende „Trace“-Anweisungen) detailliert beobachtet und beurteilt werden kann. Diese Systemfunktionen bleiben zwar theoretisch im Rahmen der Theorie der Informationsverarbeitung, da sie nur *innerhalb* dieser Theorie die Erklärbarkeit der Informationsverarbeitung ermöglichen, für andere Theorien jedoch - beispielsweise die Automatentheorie oder die Systemtheorie - können sie als nicht-theoretische Größen behandelt werden.

Inwieweit die Theorie der Informationsverarbeitung auch im Bereich geistiger Tätigkeit des Menschen, für dessen kognitive Aktivität, als grundsätzlich nicht falsifizierbar betrachtet werden soll, ist eine offene Frage. Die Methode der Computer-Simulation wird jedoch - trotz aller möglicher und tatsächlicher Verschiedenheit zwischen maschineller und menschlicher Informationsverarbeitung - bevorzugtes Instrument der kognitiven Psychologie bleiben.

## 5. Kommentiertes Literaturverzeichnis

In das Literaturverzeichnis wurde neben den im Text erwähnten Titeln eine knappe Auswahl der wichtigsten Arbeiten zu den Forschungsgebieten der Computer-Simulation und der „künstlichen Intelligenz“ (KI) aufgenommen, um einen möglichst repräsentativen Querschnitt aus der Vielfalt der bisher erschienenen Literatur zu geben. Jedem Titel ist ein kurzer kommentierender Verweis auf Inhalt und Bezug der jeweiligen Arbeit beigegeben.

## Literatur

- Abelson, R. P. 1968. Simulation of social behavior. In G. Lindzey & E. Aronson (Eds): Handbook of social psychology. Vol. 2. Reading: Addison-Wesley. - Darstellung der Simulationsmethodik in ihrer Anwendbarkeit auf sozialpsychologische Fragestellungen unter besonderer Berücksichtigung der Validierbarkeit von Simulationsmodellen.
- Anderson, J. R. 1976. Langtrage, memory, and thought. Hillsdale: Erlbaum. - Entwicklung des ACT-Simulationsmodells der menschlichen Wissensrepräsentation als kognitives Produktionssystem über einem propositionalen Netzwerk. Diskussion der Anwendbarkeit des Modells auf Inferenzprozesse, Lernen und Behalten, Verstehen und Erzeugen von Sprache und Induktion von Prozeduren.
- Anderson, J. R. & Bower, G. H. 1973. Human associative memory. Washington: Winston. - Entwurf einer Theorie der Gedächtnisrepräsentation als propositionales semantisches Netzwerk (Vorläufer der ACT-Theorie von Anderson, 1976).
- Apter, M. J. 1970. The Computer Simulation of behaviour. London: Hutchinson. - Allgemeinverständliche, inzwischen etwas veraltete Darstellung der Simulationsmethodik mit Diskussion von Anwendungen aus Bereichen des Lernens, des Problemlösens, des Mustererkennens, der Sprache und der Persönlichkeitstheorie bis hin zum Problem des Bewußtseins.
- Bauer, W. 1973. Methodische Probleme der Computer-Simulation. In G. Reinert (Hg.): Bericht über den 27. Kongreß der DGfPs in Kiel 1970. Göttingen: Hogrefe. - Einführender Artikel in die Simulationsmethodik unter dem Aspekt der Kommunizierbarkeit, Validierbarkeit und Generalisierbarkeit von Simulationsmodellen.
- Boden, M. A. 1977. Artificial intelligence and natural man. Hassocks: Harvester Press. - Ausführliche, nicht-technische Einführung in das Forschungsgebiet der KI mit einer umfassenden Darstellung der wichtigsten neueren Arbeiten aus den unterschiedlichsten Bereichen der KI-Forschung und einer eingehenden Relevanzdiskussion hinsichtlich psychologischer, philosophischer und sozialer Implikationen der KI-Forschung.
- Cohen G. 1977. The psychology of cognition. London: Academic Press. - Eine systematische, breit gefächerte Einführung in die kognitive Psychologie mit einem methodologischen Kapitel zur Computer-Simulation.
- Colby, K. M. 1975. Artificial Paranoia. A Computer Simulation of paranoid processes. New York: Pergamon Press. - Darstellung des PARRY-Programms zur Simulation paranoider Prozesse einschließlich einer Diskussion von Validierungsstudien zu dem Simulationsmodell.
- Davis, R. & King, J. 1977. An overview of production systems. In E. W. Elcock & D. Michie (Eds): Machine intelligence 8. Chichester: Horwood. - Kurzgefaßte Darstellung der Konzeption von Produktionssystemen von einem mehr technisch-methodischen Standpunkt aus.
- Dutton, J. M. & Starbuck, W. H. (Eds). 1971. Computer Simulation of human behavior. New York: Wiley. - Sammelband ausgewählter Arbeiten aus allen Be-

reichen der Computer-Simulation bis etwa zum Jahre 1970 und vollständige Bibliographie aller bis 1969 erschienenen einschlägigen Publikationen.

- Ernst, G. W. & Newell, A. 1969. GPS. A case study in generality and problem solving. New York: Academic Press. - Vollständigste Darstellung des „General Problem Solver“ als einem Simulationsmodell des menschlichen Problemlösens.
- Feigenbaum, E. A. & Feldman, J. (Eds). 1963. Computers and thought. New York: McGraw-Hill. - Erster Sammelband zu den Bereichen der Computer-Simulation und der KI-Forschung (mittlerweile von historischem Wert).
- Feurzeig, W., Lukas, G. & Grant, R. 1971. LOGO reference manual. The LOGO project NSF-C615. Cambridge, Mass.: Bolt, Beranek & Newman. - Benutzerhandbuch für die Programmiersprache LOGO in ihrer ersten Version.
- Gregg, L. W. & Simon, H. A. 1967. Process models and stochastic theories of simple concept formation. *Journal of Mathematical Psychology* 4. - Darstellung von Prozeßmodellen zur Simulation der einfachen Begriffsbildung im Vergleich zu stochastischen Theorien hierzu aus dem Bereich der mathematischen Psychologie.
- Harbordt, S. 1974. Computersimulation in den Sozialwissenschaften. 1. Einführung und Anleitung. 2. Beurteilung und Modellbeispiele. Reinbek: Rowohlt. - Allgemeinverständliche Einführung in die Simulationsmethodik von einem mehr soziologischen und weniger psychologischen Standpunkt aus.
- Heinrich, H. C. 1978. Möglichkeiten der Computersimulation in der Psychologie. *Psychologische Rundschau* 29. - überblicksartikel über die theoretischen und praktischen Möglichkeiten der Simulationsmethodik mit Diskussion von Anwendungsbeispielen.
- Hunt, E. B. 1969. Computer Simulation. *Artificial intelligence studies and their relevance to psychology. Annual Review of Psychology* 19. - Erster bibliographischer überblicksartikel über die Simulationsmethodik, ihre Grundlagen für die psychologische Modellbildung und ihre Beziehungen zur KI-Forschung.
- Hunt, E. B. & Poltrock, S. E. 1974. The mechanics of thought. In B. H. Kantowitz (Ed.): *Human information processing. Tutorials in performance and cognition*. Hillsdale: Erlbaum. - Kurzgefaßte, aber systematische Einführung in die Simulationsmethodik unter Verwendung des Produktionssystem-Ansatzes (mit Beispielen von kognitiven Produktionssystemen).
- Lenat, D. B. 1978. The ubiquity of discovery. *Artificial Intelligence* 9. - Diskussion der KI-Forschung am Beispiel des AM-Systems des Autors zum Entdecken von mathematischen Konzepten und Relationen der elementaren Zahlentheorie.
- Lenat, D. B. 1979. On automated scientific theory formation. A case study using the AM program. In J. E. Hayes, D. Michie & L. I. Mikulich (Eds): *Machine intelligence* 9. Chichester: Horwood. - Ausführliche Darstellung des AM-System des Autors mit zahlreichen Beispielen.
- McCarthy et al. 1962. LISP 1.5 programmer's manual. Cambridge, Mass.: M.I.T. Press. - Benutzerhandbuch für die Programmiersprache LISP in ihrer erstveröffentlichten Version.
- McDermott, J. & Forgy, C. 1978. Production system conflict resolution strategies. In D. A. Waterman & F. Hayes-Roth (Eds): *Pattern-directed inference systems*.

- New York: Academic Press. - Exemplarische Diskussion und Evaluation von Konfliktlösungsregeln in Produktionssystemen.
- Miller, G. A., Galanter, E. & Pribram, K. H. 1960. Plans and the structure of behavior. New York: Holt, Winston & Rinehart. (Deutsch: Strategien des Handelns. Pläne und Strukturen des Verhaltens. Stuttgart: Klett, 1973.) - Umfassender Entwurf einer neuen Psychologie auf kybernetischer Grundlage, der geradezu eine programmatische Vorwegnahme der Entwicklung der neueren kognitiven Psychologie darstellt.
- Newell, A., Shaw, J. C. & Simon, H. A. 1957. Empirical explorations with the Logic Theory Machine. A case study in heuristics. In E. A. Feigenbaum & J. Feldman (Eds): Computers and thought. New York: McGraw-Hill, 1963. - Darstellung und Diskussion eines der ersten Programme, des „Logic Theorist“, zur Simulation von Prozessen des Problemlösens, hier im Bereich der Aussagenlogik.
- Newell, A., Shaw, J. C. & Simon, H. A. 1958. Elements of a theory of human problem solving. The Psychological Review 65. - Erste, programmatische Darstellung einer Theorie des menschlichen Problemlösens auf der Grundlage der Modellierbarkeit von Denkprozessen mit Hilfe der Computer-Simulation.
- Newell, A., Shaw, J. C. & Simon, H. A. 1959. Report on a general problem solving program. In Proceedings of the International Conference on Information Processing. Paris: UNESCO House. - Bericht über das erste, als reines Simulationsmodell des menschlichen Problemlösens entwickelte Computer-Programm, den „General Problem Solver“ (GPS).
- Newell, A. & Simon, H. A. 1963. Computers in psychology. In R. D. Luce, R. R. Bush & E. Galanter (Eds): Handbook of mathematical psychology. Vol. 1. New York: Wiley. - Ausführlicher Handbuchbeitrag über den Rechnergebrauch in der Psychologie, insbesondere in Form der Computer-Simulation (mit Grundlegendiskussion und Anwendungsbeispielen).
- Newell, A. & Simon, H. A. 1972. Human problem solving. Englewood Cliffs: Prentice-Hall. - Umfassendste theoretische, methodische und empirische Darstellung des Simulationsansatzes und dessen Anwendung auf das menschliche Problemlösen in so unterschiedlichen Bereichen wie Kryptarithmetik, Logik und Schach.
- Norman, D. A., Rumelhart, D. E. & LNR Research Group. 1975. Explorations in cognition. San Francisco: Freeman. (Deutsch: Strukturen des Wissens. Stuttgart: Klett, 1978.) - Entwurf einer Repräsentationstheorie menschlichen Wissens mit Hilfe der Konzeption von semantischen Netzen und Implementierung als Computer-Modell mit Anwendungen auf Sprache, Wahrnehmung und Problemlösen.
- Ringle, M. (Ed.). 1979. Philosophical perspectives in artificial intelligence. Atlantic Highlands: Humanities Press. - Sammelband mit grundlegenden Themenstellungen und kritischen Stellungnahmen zur KI-Forschung.
- Rychener, M. D. & Newell, A. 1978. An instructable production system. Basic design issues. In D. A. Waterman & F. Hayes-Roth (Eds): Pattern-directed inference systems. New York: Academic Press. - Entwurf eines lernenden, selbstmodifizierenden Produktionssystems.
- Schank, R. C. & Abelson, R. P. 1977. Scripts, plans, goals, and understanding. An inquiry into human knowledge structures. Hillsdale: Erlbaum. - Darstellung

und Diskussion der sog. Skripttheorie menschlichen Sprachverstehens auf der Grundlage eines Computer-Modells über alltagsnahes Sprachhandeln.

- Schank, R. C. & Colby, K. M. (Eds). 1973. Computer models of thought and language. San Francisco: Freeman. - Sammelband über den Bereich der KI-Forschung in ihrer Anwendbarkeit auf Sprache und Denken.
- Simon, H. A. 1969. The sciences of the artificial. Cambridge, Mass.: M.I.T. Press. - Eine grundlegende Untersuchung der mit der Computer-Entwicklung verbundenen „Künstlichkeit“ von Wissenschaft und Welt, dargestellt an vier Themenbereichen: (1) Verstehen von natürlichen und künstlichen Welten; (2) Psychologie des Denkens: Einbettung des Künstlichen in Natur; (3) Die Wissenschaft vom Entwerfen: Erzeugung des Künstlichen; (4) Die Architektur von Komplexität.
- Simon, H. A. 1979. Information processing models of cognition. Annual Review of Psychology 30. - Bibliographische Darstellung des Ansatzes der Informationsverarbeitung unter besonderer Berücksichtigung der Simulationsmethodik.
- Simon, H. A. 1979a. Models of thought. New Haven: Yale University Press. - Sammelband über eine repräsentative Auswahl von Arbeiten des Autors und seiner Mitarbeiter zur kognitiven Psychologie von 1955 bis 1977.
- Sloman, A. 1978. The Computer revolution in philosophy. Philosophy, science, and models of mind. Hassocks: Harvester Press. - Eine in philosophische und wissenschaftstheoretische Fragestellungen des Rechnergebrauchs in der KI-Forschung und in der kognitiven Psychologie eingehende Untersuchung der Computer-Metapher.
- Stegmüller, W. 1973. Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band 2: Theorie und Erfahrung. Studienausgabe Teil D: Logische Analyse der Struktur ausgereifter physikalischer Theorien. „Non-statement view“ von Theorien. Studienausgabe Teil E: Theoriendynamik. Normale Wissenschaft und wissenschaftliche Revolutionen. Methodologie der Forschungsprogramme oder epistemologische Anarchie? Berlin: Springer. - Detaillierte Darstellung der strukturalistischen Theoriekonzeption („non-statement view“) am Beispiel physikalischer Theorienbildungen und wissenschaftstheoretischer Paradigmenvorstellungen.
- Stegmüller, W. 1980. Neue Wege der Wissenschaftsphilosophie. Berlin: Springer. - Weiterführende Arbeiten zur strukturalistischen Theoriekonzeption („non-statement view“).
- Tomkins, S. S. & Messick, S. (Eds). 1963. Computer Simulation of personality. New York: Wiley. - Sammelband über erste Arbeiten zur Anwendung der Simulationsmethodik auf Bereiche der Persönlichkeitsforschung.
- Turing, A. M. 1950. Computing machinery and intelligence. Mind 59. (Reprinted in E. A. Feigenbaum & J. Feldman (Eds): Computers and thought. New York: McGraw-Hill, 1963. Deutsch: Kann eine Maschine denken? Kursbuch 8, 1967.) - Berühmte Arbeit zu der Frage, ob Computer „denken“ können, mit einer Diskussion einer Reihe grundsätzlicher Gegenargumente.
- Ueckert, H. 1980a. Cognitive production systems. Toward a comprehensive theory of mental functioning. In F. Klix & J. Hoffmann (Eds): Cognition and memory. Knowledge and meaning comprehension as functions of memory. Amsterdam:

North-Holland Publishing Company. - Diskussion der Produktionssystem-Konzeption als einer umfassenden Theorie kognitiver Aktivität.

- Ueckert, H. 1980b. The cognitive executive. From artificial intelligence toward a psychological theory of consciousness. In Proceedings of the XXIIInd International Congress of Psychology in Leipzig. - Anwendung der Produktionssystem-Konzeption auf die Bewußtseinsproblematik der menschlichen Informationsverarbeitung.
- Ueckert, H. 1980c. Das Lösen von Intelligenztestaufgaben. Eine test- und meßkritische Untersuchung. Göttingen: Hogrefe. - Entwicklung eines prozeßorientierten Ansatzes zur Intelligenzforschung auf der Grundlage der Produktionssystem-Konzeption und Diskussion der Axiomatisierbarkeit der Intelligenztestung als additiv-verbundener Messung.
- Ueckert, H. & Rhenius, D. (Hg.). 1979. Komplexe menschliche Informationsverarbeitung. Beiträge zur Tagung „Kognitive Psychologie“ in Hamburg 1978. Bern: Huber. - Bericht über die erste Fachtagung zur kognitiven Psychologie in der BRD mit sechs Themenschwerpunkten: (1) Kognitive Psychologie als integrativer Bestandteil psychologischer Grundlagenforschung; (2) Kognitive Organisation der menschlichen Informationsverarbeitung; (3) Informationelle Produktionssysteme und Computer-Simulation; (4) Sprachliche Kognition und semantisches Gedächtnis; (5) Entscheidungstheoretische Ansätze zur kognitiven Psychologie; (6) Anwendungsfragen der kognitiven Psychologie.
- Uhr, L. 1973. Pattern recognition, learning, and thought. Computer-programmed models of higher mental processes. Englewood Cliffs: Prentice-Hall. - Eine ganz auf die methodischen Fertigkeiten des Programmierens (in der hierzu vom Autor konzipierten Programmiersprache EASEY-1) angelegte Einführung in den Bereich des Computer-gestützten Modellierens von Prozessen des Wahrnehmens, Lernens und Denkens.
- Vukovich, A. F. 1967. Die Simulierung des Problemlösens auf logischen Automaten. In F. Merz (Hg.): Bericht über den 25. Kongreß der DGfPs in Münster 1966. Göttingen: Hogrefe. - Erster deutschsprachiger überblicksartikel über die Computer-Simulation in der psychologischen Forschung unter theoretischen und methodologischen Aspekten.
- Waterman, D. A. & Hayes-Roth, F. (Eds). 1978. Pattern-directed inference systems. New York: Academic Press. - Sammelband mit einer systematischen Auswahl von Arbeiten auf der Grundlage der Produktionssystem-Konzeption mit Anwendungen im Bereich der Computer-Simulation und in der KI-Forschung.
- Wegener, H. & Dörner, D. 1973. Simulation als Forschungstechnik. Bericht über ein Symposium. In G. Reinert (Hg.): Bericht über den 27. Kongreß der DGfPs in Kiel 1970. Göttingen: Hogrefe. - Zusammenfassende Darstellung einer Diskussion über die Computer-Simulation als Forschungsinstrument in der Psychologie.
- Weizenbaum, J. 1976. Computer power and human reason. From judgment to calculation. San Francisco: Freeman. (Deutsch: Die Macht der Computer und die Ohnmacht der Vernunft. Frankfurt am Main: Suhrkamp, 1977.) - Eine sehr kritische, kompetente Diskussion der Computer-Metapher und ihrer Verwendung in Forschung und Praxis.

- Wender, K. F., Colonius, H. & Schulze, H.-H. 1980. Modelle des menschlichen Gedächtnisses. Stuttgart: Kohlhammer. - Kurzgefaßte Darstellung der wichtigsten Ansätze zur Gedächtnisrepräsentation auf der Grundlage von semantischen Netzen.
- Wexler, K. 1978. A review of John R. Anderson's *Language, Memory, and Thought*. *Cognition* 6. - Ausführliche und kritische Besprechung der ACT-Theorie von Anderson (1976).
- Winograd, T. 1972. Understanding natural language. *Cognitive Psychology*, Whole Number 3. (Also published by Academic Press, New York.) - Darstellung eines der ersten KI-Systeme zum natürlichsprachlichen Verstehen auf der Grundlage einer Computer-simulierten „Mini-Welt“.
- Winston, P. H. 1977. Artificial intelligence. Reading: Addison-Wesley. - Eine sehr instruktiv geschriebene und illustrative Einführung in den Forschungsbereich der „künstlichen Intelligenz“ mit besonderer Berücksichtigung des Programmierens (in der Programmiersprache LISP).

# Autoren-Register

Hinweis: Die *kursivgedruckten* Seitenangaben beziehen sich auf die Literaturverzeichnisse der Artikel.

- Abelson, R. P. 563, 596, 611, 613  
 Abrahams, N. M. 91, 119, 172, 192  
 Abrami, P. F. 182, 220, 223  
 Acock, A. C. 167f, 192, 232  
 Adam, J. 93, 192  
 Adams, E. R. 40, 192  
 Afsarinejad, K. 133, 221  
 Ahtola, O. T. 130, 237  
 Aitken, A. C. 297, 453  
 Akaike, H. 419, 453  
 Alf, E. F. Jr. 91, 119, 172, 192  
 Algina, J. 315, 318, 325, 362, 453, 468  
 Alimena, B. S. 57, 192  
 Allen, D. M. 315, 320, 325, 363, 459  
 Allerbeck, K. R. 40, 192  
 Alwin, D. F. 343, 469  
 Amemiya, T. 370, 453  
 Amir, Y. 47, 192  
 Anastasio, E. J. 130, 204  
 Anderson, B. L. 125, 213  
 Anderson, J. R. 563, 574, 593f, 611  
 Anderson, N. H. 40, 192  
 Anderson, O. D. 266, 453  
 Anderson, T. W. 353, 357, 408ff, 415, 453  
 Andreski, S. 421, 453  
 Andress, H. J. 418, 453  
 Andrews, D. F. 117f, 192  
 Anscombe, F. J. 116, 192  
 Appelbaum, M. I. 20, 22, 140, 143, 182, 201, 221  
 Applebaum, M. I. 295, 463  
 Armenakis, A. A. 121, 204  
 Arminger, G. 436, 438, 453  
 Armitage, P. 32, 193  
 Arnold, W. 50, 193  
 Aronson, E. 27, 36f, 44f, 50, 59, 66, 193, 198  
 Asad, H. 112, 229  
 Aschenbrenner, K. M. 509f, 519, 526  
 Aspin, A. A. 113, 193  
 Astrom, K. J. 438, 454  
 Bailey, D. E. 174, 187, 193  
 Bakan, D. 81, 83f, 193  
 Baker, B. O. 40, 193  
 Baker, F. B. 91, 141, 193, 201  
 Balestra, P. 370, 454  
 Baltes, P. B. 239, 243, 454, 465  
 Baran, S. J. 85, 199  
 Barber, T. X. 32, 43, 193  
 Barlow, D. H. 67, 212, 285, 454, 460  
 Barlow, R. E. 127, 193  
 Barnett, V. 70f, 118, 193  
 Barr, D. R. 130, 198  
 Bartenwerfer, H. 193  
 Bartholomew, D. J. 127, 193, 194, 243, 418, 454  
 Bartlett, M. S. 112, 114, 142, 182, 194, 277, 367, 410, 454  
 Battese 373, 375, 458  
 Bauer, W. 230  
 Bauknecht, K. 165, 194  
 Bayes, T. 478, 500, 526  
 Beauchamp, J. J. 110, 195  
 Beauchamp, K. B. 84, 194  
 Beauchamp, K. L. 50, 221  
 Becker, P. 328, 454  
 Behnken, D. W. 116, 194  
 Behrens, W. V. 113, 194  
 Bellmann, R. 438, 454  
 Bern, D. J. 42, 194  
 Benninghaus, H. 194  
 Benton, J. Q. 110, 194  
 Berchthold, H. 93, 96, 194  
 Bereiter, C. 240ff, 330, 332, 343, 454, 469  
 Berenson, M. L. 127, 194  
 Berg, J. A. 43, 194  
 Berger, P. K. 58, 224  
 Bergstrom, A. R. 441, 454  
 Bernitzke, F. 4, 5, 22  
 Bernhardson, C. S. 194



- Betz, M. A. 145, 194  
 Bevan, M. F. 110, 194  
 Bickel, P. J. 117f, 192, 194  
 Bisbee, C. J. 130, 211  
 Bishop, T. A. 176, 194  
 Blackwell, D. 503, 526  
 Blair, R. C. 95, 110, 114, 181, 194, 195  
 Blalock, H. M. 59, 94, 107f, 114, 223, 422, 454, 465  
 Blankenship, V. 422, 462  
 Bliss, C. J. 94, 104, 106f, 116f, 195, 200  
 Bloomfield, P. 354, 454  
 Blumen, J. 414, 416, 454  
 Bock, J. 73, 94, 105, 170f, 187, 227  
 Bock, R. D. 94, 195, 240, 296, 333, 335f, 339, 344, 349, 351, 356f, 359, 361, 365, 367, 369, 443, 455  
 Bock, R. R. 94, 201  
 Boden, M. A. 580, 611  
 Boersma, F. D. 127, 195  
 Boes, D. C. 72, 223  
 Boik, R. J. 125, 195  
 Bolles, R. C. 158, 163, 195  
 Boneau, C. A. 114, 195  
 Borcharding, K. 509, 526  
 Borich, G. O. 149, 195  
 Bortz, J. 72, 77, 82, 84f, 88, 94, 107f, 116, 133, 140, 164ff, 183, 189, 195, 287, 455  
 Bower, C. P. 455  
 Bower, G. H. 21, 22, 406, 455, 563, 611  
 Bower, S. M. 287, 467  
 Bowman, K. O. 110, 195  
 Box, G. E. P. 94, 106, 109, 111f, 114, 116, 136, 140f, 195f, 245, 254, 259, 262, 264ff, 268, 273, 277, 282, 295, 297, 304, 348, 367, 441, 455, 496, 524, 526  
 Boyett, J. M. 124, 231  
 Bozarth, J. D. 71, 84, 196  
 Bracht, G. H. 60, 63, 147, 150f, 154, 196, 286, 455  
 Bradley, J. V. 90, 93, 95f, 110, 119, 196  
 Brand, M. 28, 196  
 Brandis, H. P. 125, 202  
 Brandtstädter, J. 4f, 8, 22  
 Braskamp, L. A. 40, 211  
 Bredenkamp, J. 3, 5, 8, 14f, 18, 22, 26, 29f, 32, 36, 44, 51, 55f, 58f, 62, 66, 68f, 71f, 78, 81, 84f, 91, 93f, 97f, 102, 108, 116, 128f, 133, 142, 145, 147, 152, 154f, 158, 161, 163ff, 169, 172, 180ff, 184, 187ff, 192, 196, 197, 475, 482, 491, 499, 526  
 Breen, L. 125, 141, 163, 197, 216, 228  
 Bremner, J. M. 127, 193  
 Brewer, J. K. 85, 187, 197  
 Bridgeman, P. W. 38, 197  
 Brillinger, D. R. 136, 197  
 Brocke, B. 66, 197  
 Broekmann, N. C. 116, 197  
 Brooks, W. D. 94, 229  
 Brown, B. M. 112, 219  
 Brown, D. J. 106, 197  
 Brown, G. D. 285, 469  
 Brown, M. B. 111ff, 197, 198  
 Bruce, R. L. 50, 221  
 Brunk, H. D. 127, 193  
 Büning, H. 93, 95f, 109f, 198  
 Bugelski, B. R. 50, 198  
 Bungard, W. 43, 66, 198  
 Bunge, M. 28f, 38, 198  
 Burke, C. J. 243, 395, 462  
 Burnett, T. D. 130, 198  
 Busch, K. 73, 94, 105, 170f, 187, 227  
 Busemeyer, J. R. 152, 198  
 Bush, N. 145, 204  
 Bush, R. R. 174, 198, 223  
 Byatt, S. E. 30, 235  
 Cadzow, J. A. 424, 439, 455  
 Callaway, J. N. 43, 198  
 Campbell, D. T. 1, 22, 27ff, 32, 35, 47f, 54, 58ff, 65f, 70, 136, 198, 201, 236, 240, 284, 287, 455, 456  
 Carlsmith, J. M. 27, 36f, 44f, 50, 59, 66, 161, 193, 198, 205  
 Carlson, R. 72, 198  
 Carmer, S. G. 124, 198  
 Carnap, R. 33, 35, 199, 486, 526  
 Carroll, R. J. 118, 236  
 Carroll, R. M. 165, 199  
 Carter, D. S. 164f, 199  
 Carter, L. F. 59, 94, 107f, 114, 223, 456  
 Carver, R. P. 72, 199  
 Cascio, W. F. 85, 199  
 Cattell, R. B. 240, 456  
 Chan, S. G. 423, 430, 442, 456  
 Chan, S. P. 423, 430, 442, 456  
 Chan, S. Y. 423, 430, 442, 456  
 Chardos, S. 115, 220

- Chase, L. J. 72, 85, 199, 217  
 Chase, R. B. 85, 199  
 Chassan, J. B. 243, 285, 456  
 Chatfield, C. 136, 199, 265, 456  
 Chen, H. J. 109, 231  
 Chomsky, N. 395, 464  
 Chow, G. C. 385, 456  
 Church, J. S. 112f, 124, 199, 237  
 Clauss, G. 113, 199, 328, 456  
 Cleary, T. A. 172, 199, 200  
 Clinch, J. J. 112f, 208  
 Cochran, W. G. 32, 56f, 73, 94, 104, 106, 108, 110ff, 130f, 172, 174, 200, 232  
 Cohen, J. 14, 22, 71, 84f, 94, 104, 124, 137, 141, 158, 160ff, 168, 174, 177ff, 200, 482, 490ff, 526  
 Cohen P. 94, 124, 137, 141, 162ff, 177, 180, 182, 200  
 Colby, K. M. 597, 611, 614  
 Coleman, J. S. 330, 395, 414, 418, 436, 438, 456  
 Collier, R. O. J. 91, 141, 165, 193, 201  
 Colonius, H. 563, 598, 616  
 Conlisk, J. 414, 416, 456  
 Conover, W. J. 110, 114, 215  
 Cook, S. W. 35, 43, 201  
 Cook, T. D. 27ff, 32, 47, 48, 54, 62, 65, 70, 136, 201, 236, 284, 456, 486, 526  
 Cooley, W. W. 94, 162, 201, 262, 456  
 Coombs, C. H. 98, 201, 474, 526  
 Cooper, H. M. 30, 201  
 Corballis, M. C. 158, 165, 235  
 Cowles, M. P. 83, 185, 201  
 Cox, D. R. 57, 94, 108, 114, 131, 172, 196, 201, 406f, 417, 456  
 Cox, G. M. 56f, 94, 174, 200  
 Craig, A. T. 212  
 Craig, J. R. 201  
 Crain, B. R. 139f, 222  
 Cramer, E. M. 94, 163, 182, 201  
 Cronbach, L. J. 35, 136, 149, 201, 202, 332, 456  
 Cureton, E. E. 165, 202  
 D'Agostino, R. B. 93, 202  
 Dalal, S. N. 172, 225  
 Danks, J. H. 168, 203  
 Darlington, R. B. 104, 202  
 Da Silva, J. G. C. 374, 456  
 Davenport, S. M. 440, 467  
 David, D. J. 202  
 David, H. A. 125, 202  
 Davidson, M. L. 138, 141, 202  
 Davis, D. J. 43, 44, 126, 229  
 Davis, F. 299, 459  
 Davis, R. 578, 611  
 Dawes, R. M. 40, 98, 201, 202  
 Dayton, C. M. 124, 202  
 De Friesse, F. 124, 234  
 De Jonge, J. J. 127, 195  
 Dénes, J. 57, 202  
 Deppe, W. 15, 20f, 22, 395, 421, 456  
 Derrick, T. 72, 202  
 Dertouzos, M. L. 394, 423, 427, 438, 453  
 Desu, M. M. 110, 114, 218  
 Deutsch, S. J. 419, 465  
 Diananda, P. H. 454  
 Diehl, J. M. 94, 125, 144, 174, 175, 177, 202  
 Dierkes, M. 136, 202  
 Digman, J. M. 114, 147, 202  
 Dipboye, R. L. 65, 202  
 Dippner, R. S. 110, 234  
 Dixon, W. J. 118, 172, 174f, 202  
 Dodd, D. H. 165, 203  
 Dooling, D. J. 168, 203  
 Doreian, P. 425, 426, 436, 438, 457, 461  
 Draper, N. R. 94, 114, 116, 118, 194, 203  
 Drösler, J. 385, 457  
 Dubins, L. 503, 526  
 Duke, M. P. 43, 198  
 Dunlap, W. P. 84, 85, 218  
 Dunn, O. J. 121, 124f, 203  
 Dunnett, C. W. 121, 122, 124f, 203  
 Dutton, J. M. 535, 611  
 Dwyer, J. H. 165, 203  
 Dyer, J. S. 519, 526  
 Ebenhöf, W. 421, 457  
 Ebnet-, H. 113, 199  
 Edgington, E. S. 71, 90f, 203, 287, 457  
 Edwards, A. L. 43, 48, 54, 57, 88, 94, 107, 158, 164, 203  
 Edwards, A. W. F. 71, 204  
 Edwards, W. 70, 204, 503, 518, 524f, 526, 527  
 Eimer, E. 94, 141, 204

- Einot, J. 125, 204  
 Eisenhart, C. 104, 204  
 Eison, C. L. 201  
 Ekbohm, G. 111, 113, 204  
 Elashoff, J. D. 130, 204, 287, 292, 457, 468  
 Elashoff, R. M. 108, 204  
 Ellsworth, P. C. 36, 37, 44f, 50, 59, 66, 198  
 Elston, R. C. 107, 145, 204  
 Emerson, M. 299, 459  
 Enderlein, G. 108, 113, 136, 170f, 174, 204, 227  
 Engelhardt, W. 145, 204  
 Enke, H. 93, 192  
 Erlebacher, A. 136, 204  
 Estes, W. K. 395, 454, 457  
 Evans, S. H. 130, 204, 211  
 Everitt, B. S. 142, 204  
  
 Fagot, R. F. 40, 192  
 Fararo, T. J. 395, 457  
 Federer, W. T. 130, 204, 361, 470  
 Feger, H. 85, 197  
 Feigenbaum, E. A. 608, 612  
 Feild, H. S. 121, 204  
 Feir-Walsh, B. E. 110f, 114, 204  
 Feir-Walsh, B. J. 111, 114, 216  
 Feldmann, J. 608, 612  
 Feldt, L. S. 129, 131, 140f, 175, 204, 205, 214, 237, 348, 461  
 Fennessey, J. 103f, 106, 205  
 Ferguson, G. A. 93, 205  
 Ferschl, F. 405ff, 457  
 Festinger, L. 27, 42, 161, 201, Feuerzeig, W. 551, 612  
 Fichter, M. M. 285, 457  
 Fietkau, H.-J. 26, 205  
 Fieve, R. R. 259, 461  
 Finn, J. D. 94, 205, 296, 351, 357, 365, 458  
 Finetti, B. de 478, 509, 527  
 Firth, J. 163, 207  
 Fischer, G. 43, 107, 205, 420, 449, 458  
 Fischer, G. H. 243, 458  
 Fisher, R. A. 54, 56, 71f, 76, 79, 82, 84, 85, 90, 104, 107, 113, 117, 124, 161, 163, 205, 287, 458, 476, 495, 527  
 Fiske, D. W. 35, 198  
 Fisz, M. 72, 73, 119, 205, 331, 458, 484, 497, 527  
 Fitting, U. 395, 466  
 Flanagan, M. F. 65, 202  
 Flay, B. R. 486, 526  
 Fleishman, A. J. 162, 165, 168, 205, 213  
 Fleiss, J. L. 92, 141, 164f, 168, 172, 183, 205, 206, 236  
 Forge, C. 575, 577f, 612  
 Forsyth, R. A. 172, 206  
 Forsythe, A. B. 11 1ff, 197, 198  
 Fox, M. 163, 174, 177, 224  
 Fox, R. 299, 459  
 Frame, J. S. 439, 458  
 Fredericksen, S. H. 92, 222  
 French, J. R. P. 59, 206  
 Frey, D. 27, 206  
 Fricke, R. 30, 206  
 Friedman, H. 163, 206  
 Friedman, M. 93, 145, 167, 206  
 Fromkin, H. L. 50, 206  
 Fruchter, B. 127, 210  
 Fuller, W. A. 353f, 373, 375, 458  
 Furby, L. 136, 202, 332, 456  
  
 Gabriel, K. R. 125, 145, 194, 204, 206  
 Gabriel, R. M. 133, 206  
 Gadenne, V. 27ff, 32f, 60, 62f, 66f, 69, 206  
 Gaebelien, J. W. 163, 165, 182, 207, 212  
 Gaensslen, H. 22, 82, 94, 106, 109, 182, 207  
 Gaito, J. 40, 84, 89, 94f, 105f, 110, 124, 141, 163, 165, 197, 207, 228, 240, 344, 458  
 Garmes, P. A. 111ff, 121, 124ff, 145, 207, 208, 213, 215, 217, 221  
 Gardner, P. L. 40, 208  
 Garten, H.-K. 149, 208  
 Gartside, P. S. 112f, 208  
 Gastwirth, J. L. 296, 458  
 Geary, R. C. 108f, 208  
 Gebert, A. 145, 208  
 Gebhardt, F. 108, 208  
 Gehan, E. A. 110, 114, 218  
 Geisser, S. 141, 208, 209, 295, 344, 348, 458, 459  
 Gentile, J. R. 287, 458  
 Gentle, J. E. 112, 234  
 Giambalvo, V. 110, 234  
 Gibbons, J. D. 89, 93, 208  
 Guillo, M. W. 163, 231  
 Ginsberg, R. B. 414f, 418, 458  
 Glaser, W. R. 15f, 22  
 Glasnapp, D. R. 165, 168, 210  
 Glass, G. V. 30, 54, 60, 63, 89, 105ff, 109ff, 114ff, 130, 136, 147, 150f, 153f, 164, 167f, 181f, 192, 296, 208, 209, 214, 244f, 250, 252, 254,

- 266, 286f, 296f, 299,  
301, 303f, 455, 458,  
459
- Glavin, G. B. 133, 206
- Gniech, G. 43, 209
- Godbout, R. C. 149,  
195
- Göricke, G. 245, 247,  
249, 305, 464
- Gokhale, D. V. 93, 209
- Gold, D. 158, 209
- Goldberger, A. S. 339,  
459
- Goldsmith, L. 299, 459
- Goldstein, H. 341, 420,  
459
- Golhar, M. B. 113, 209
- Goodman, L. A. 92,  
209, 410, 412, 416f, 453,  
459
- Gottman, J. M. 136,  
209, 243ff, 250, 252,  
254, 266, 287, 296f,  
299, 301, 303f, 459
- Granger, C. W. J. 114,  
223, 282, 459
- Grant, R. 612
- Gray, H. L. 118, 209
- Graybill, F. A. 72, 107,  
214, 223
- Greenhouse, S. W. 141,  
208, 209, 295, 344, 348,  
458, 459
- Greeno, J. G. 98, 227,  
395, 397, 406, 438, 459,  
466
- Greenwald, A. G. 48ff,  
58, 84, 135, 141, 209
- Gregg, L. W. 536, 538,  
545, 599, 612
- Grizzle, J. 315, 320,  
325, 363, 459
- Grizzle, J. E. 92, 209
- Groebe, N. 28, 61, 69,  
209
- Gröbner, W. 437, 459
- Groenveld, L. P. 418,  
468
- Groot, M. H. de 507,  
524, 527
- Gruder, C. L. 486, 526
- Gruijter, D. N. M. de  
136, 209
- Guenther, A. L. 84, 231
- Guenther, W. C. 174,  
176f, 209
- Guilford, J. P. 127, 210
- Guthke, J. 328, 456, 459
- Guttentag, M. 518, 526
- Guttman, L. 72, 85f,  
161, 210
- Haagen, K. 72f, 77f, 82,  
84f, 117, 171, 210
- Haberman, S. J. 210
- Hacking, J. 72, 210
- Hagen, K. 511, 527
- Hager, W. 29, 32, 81,  
95, 97ff, 110ff, 121f,  
127, 144, 155, 161f,  
189ff, 197, 210, 237
- Haggard, E. A. 94, 195
- Hakstian, A. R. 142,  
168, 209, 210
- Halderson, J. S. 165,  
168, 210
- Hall, B. H. 370, 459
- Hall, J. J. 112f, 210
- Hall, R. V. 299, 459
- Hamer, R. M. 181, 210
- Hamilton, B. L. 130,  
210
- Hammersley, J. M. 165,  
211
- Hamouzova, M. 387,  
460
- Hampel, F. R. 116ff,  
192, 211
- Handscorn, D. C. 165,  
211
- Hanit, M. 112, 229
- Hannan, E. J. 419, 460
- Hannan, M. T. 243, 375,  
418, 460, 468
- Harbordt, S. 533f, 550,  
554f, 612
- Harder, Th. 395, 460
- Hardyck, C. D. 40, 120,  
193, 226
- Harnatt, J. 72, 84f, 211
- Harris, C. W. 136, 211
- Harris, Ch. W. 240, 460
- Harris, D. B. 130, 211
- Harris, R. J. 94, 211
- Hart, M. C. 438, 460
- Harter, H. L. 125, 211
- Hartley, H. O. 54, 84,  
112f, 124, 174, 211, 225
- Hartmann, D. P. 287,  
460
- Havlicek, L. L. 110ff,  
211
- Hawkins, D. M. 118,  
211
- Hay, R. A. 259, 264,  
277, 311, 463
- Hayes, T. F. 141, 201
- Hayes-Roth, F. 564,  
579f, 615
- Hays, W. L. 71ff, 77,  
79, 81ff, 88f, 104, 108,  
111, 115f, 120, 122,  
124, 144, 158, 161f,  
165ff, 171, 180f, 191,  
211, 491, 507, 527
- Heckhausen, H. 50, 211
- Hedayat, A. 133, 211
- Heermann, E. F. 40, 211
- Hegemann, V. 107, 211
- Hehl, F. J. 67, 136, 226,  
239, 465
- Helmer, R. M. 244, 278,  
460
- Hempel, C. G. 33, 35,  
61, 211
- Henderson, C. R. 375,  
460
- Henkel, R. E. 72, 223
- Henning, H. J. 28, 32,  
94, 106, 108, 114f, 144,  
149, 154, 164, 181, 211,  
482, 490, 527

- Henningan, K. 486, 526  
 Henry, N. W. 417, 460, 462  
 Henze, F. H.-H. 184, 212  
 Herr, D. G. 181, 212  
 Herrendörfer, G. 73, 94, 105, 108, 113, 136, 170, 171, 174, 187, 227  
 Herrmann, T. 21, 23, 33, 38, 41, 68, 108, 212  
 Hersen, M. 67, 212, 285, 454, 460  
 Hibbs, D. A. 287, 296, 297, 460  
 Hibbs, D. A. Jr. 249, 255, 313, 460  
 Higbee, K. L. 86, 212  
 Hesse, H. G. 103, 104, 193  
 Higgins, J. J. 95, 110, 114, 181, 194, 195  
 Hilgard, J. R. 304f, 460  
 Hochberg, Y. 125, 176, 212  
 Hodges, J. L. 95f, 158, 212  
 Hofstätter, P. R. 503, 506, 512, 527  
 Hogg, R. V. 118, 212  
 Hohander, M. 93, 212  
 Hollingsworth, H. H. 130, 212  
 Holm, K. 43, 212  
 Holtzman, W. H. 240, 461  
 Holzkamp, K. 63, 212  
 Hopkins, K. D. 125, 133, 165, 206, 212, 213  
 Horan, P. H. 415, 461  
 Horsnell, G. 112, 213  
 Horst, P. 240, 461  
 Hosking, J. D. 181, 210  
 Hotelling, H. 183, 213  
 Howard, R. A. 418, 438, 461  
 Howell, J. F. 111f, 125, 208, 213  
 Hoyle, M. H. 114, 213  
 Hsu, L. M. 121, 213  
 Huba, G. J. 259, 461  
 Huber, P. J. 117f, 192, 213  
 Hubert, L. J. 94, 213, 228, 285, 287, 289, 292, 462  
 Huberty, C. J. 162, 164f, 213  
 Huck, S. W. 58, 153, 213  
 Hübner, R. 110, 113, 210, 213  
 Huitema, B. E. 130, 213  
 Hummel, T. J. 143, 182, 213  
 Hummell, H. J. 3f, 19, 23, 59, 213  
 Hummon, N. P. 425f, 436, 438, 457, 461  
 Humphreys, L. G. 162, 213  
 Hunt, E. B. 563, 612  
 Hunter, J. E. 85, 230  
 Hunter, J. S. 94, 196  
 Hunter, W. G. 114, 203  
 Hunter, W. G. Jr. 94, 196  
 Hurlburt, R. T. 121, 144, 214  
 Hussain, A. 370, 375, 469  
 Hussian, R. A. 85, 163, 168, 232  
 Huyngh, H. 20, 23, 139f, 214, 348, 461  
 Irle, M. 27, 214  
 Isaak, P. D. 342, 461  
 Jackson, P. H. 524, 528  
 Jacobs, K. W. 120, 214  
 Jäger, R. 108, 230  
 James, G. S. 113, 214, 235  
 Jenkins, G. M. 136, 196, 245, 249, 252, 254f, 259, 262, 264f, 268, 272f, 277, 281f, 348, 455, 461  
 Jennings, E. 104, 182, 214  
 Jennrich, R. J. 440, 466  
 Jöreskog, K. G. 326, 342f, 347f, 359, 374f, 378, 382, 390, 392, 423, 433, 461, 469  
 Johannsson, J. K. 244, 278, 460  
 John, J. A. 56f, 94, 106, 214  
 John, P. W. M. 106, 181, 214  
 Johnson, D. E. 107, 211, 214  
 Jonckheere, A. R. 127, 214  
 Judge, G. G. 243, 410, 462  
 Jurs, S. G. 182, 214  
 Kaiser, H. F. 89, 189, 214, 240, 461  
 Kamp, L. J. T. van der 136, 209  
 Katz, B. M. 90, 95, 222  
 Katzer, J. 85, 214  
 Kay, K. J. 30, 235  
 Kazdin, A. E. 50, 215, 285, 288, 461  
 Keedwell, A. D. 57, 202  
 Keeney, R. L. 514, 527  
 Keeser, W. 244, 266, 275, 296, 298, 300, 442, 466  
 Kelly, F. J. 296, 464  
 Kemeny, J. G. 406, 461  
 Kemmnitz, W. 510, 525, 527  
 Kemp, K. E. 110, 114, 215  
 Kempthorne, O. 90, 94, 104, 107, 215, 237

- Kendall, M. G. 72, 73, 89, 136, 158, 167, 215, 335, 341f, 461, 470
- Kennedy, J. J. 162, 168, 215
- Kenny, D. A. 59, 215
- Keppel, G. 48, 94, 115f, 125, 138, 141, 176f, 215
- Kerchner, M. 30, 235
- Keren, G. 162, 181, 215, 220
- Kerlinger, F. N. 5, 13, 23, 59f, 94, 144, 161f, 180, 182, 215
- Keselman, H. J. 111, 114, 124f, 139f, 165, 215, 216, 228, 287, 462
- Keselman, H. K. 112f, 208
- Khatri, C. G. 315, 326, 344, 365, 384, 462
- Kimball, A. W. 120, 144, 216
- King, J. 578, 611
- Kirk, R. E. 40, 94, 122, 124f, 130, 133, 137, 140, 159, 174f, 177, 216
- Klauer, K. J. 36, 43, 63, 216
- Kleijnen, J. P. C. 165, 216
- Klein, R. D. 287, 458
- Kleiter, E. 239, 462
- Kleiter, G. 72, 83, 216
- Kleiter, G. D. 70, 189, 216
- Klett, C. J. 94, 225
- Knapp, T. R. 104, 142, 216
- Koch, G. 92, 209
- Koch, J. J. 43, 216
- Kogan, M. 414, 416, 454
- Kohlas, J. 165, 194
- Kohr, R. L. 111f, 125f, 217
- Kolmogoroff, A. N. 477, 527
- Konijn, H. S. 217
- Konkin, P. R. 127, 221
- Koran, J. J. Jr. 422, 463
- Kormann, A. 239, 462
- Kraemer, H. C. 184, 217
- Krantz, D. H. 40, 217
- Kranz, H. T. 43, 107, 217
- Krapp, A. 239, 462
- Kratochwill, T. R. 67, 217
- Krause, B. 81, 84, 187, 217
- Krauth, J. 93, 217
- Kreyszig, E. 54, 217
- Kriz, J. 54, 72, 217
- Kröh, P. 245, 247, 249, 305, 464
- Kroll, R. M. 85, 217
- Krüger, H.-P. 145, 217
- Kruglanski, A. W. 43, 217
- Kruskal, W. H. 93, 145, 166, 217
- Ku, H. H. 93, 217
- Küchler, M. 92, 217
- Küttner, M. 61, 217
- Kuhl, J. 422, 462
- Kullback, S. 93, 209, 217
- Kuo, F. K. 440, 466
- Kupst, M. J. 84, 230
- Kuriakjian, B. 462
- Kutschera, M. F. von 486, 527
- Labovitz, S. 84, 218
- Lachenbruch, P. A. 125, 176, 202, 212
- Läuter, J. 163, 183, 218
- La Forge, R. 218
- Lakatos, J. 30, 33, 218
- Laming, D. 395, 418, 462
- Lana, R. E. 62, 116, 218
- Land, K. C. 436, 438, 462
- Lane, D. M. 84f, 218
- Langcheine, R. 92, 168, 218
- Lantermann, E. D. 40, 218, 239, 467
- Larson, R. C. 165, 201
- Lawler, W. G. 259, 461
- Layard, M. W. J. 112, 218
- Lazarsfeld, O. 417, 462
- Lee, E. T. 110, 114, 218
- Lee, M. C. 106, 147, 218
- Lee, T. C. 243, 410, 462
- Lee, W. 94, 164, 174, 218
- Leeb, S. 172, 218
- Lehmann, G. 12, 23
- Lehman, E. L. 93, 95f, 111, 158, 181, 187, 189, 212, 218, 288, 462
- Lehwald, G. 328, 456
- Leiblum, S. R. 243, 459
- Leiser, E. 72, 82, 219
- Lenat, D. B. 586, 612
- Lépine, D. 139f, 229
- Leslie, R. I. 112, 219
- Levene, H. 112, 219
- Leventhal, L. 287, 462
- Levin, J. R. 143, 145, 172, 177, 219, 221, 228, 233, 243, 285, 287, 289, 292, 462
- Levine, G. 395, 462
- Levine, M. 21, 23
- Levy, K. J. 112f, 124, 182, 219, 220, 223
- Levy, P. 158, 163, 220
- Lewandowski, R. 419, 462
- Lewis, C. 162, 181, 215, 220
- Lewis, T. 118, 193
- Li, C. C. 113, 220
- Lichtenstein, S. 524, 528
- Lieberman, B. 40, 89, 220
- Lienert, G. A. 19, 23, 90f, 95f, 114, 127, 136, 143, 145, 158, 161, 166f, 217, 220, 498, 511, 527

- Light, R. J. 30, 192, 226  
 Lind, J. C. 142, 210  
 Lindley, D. V. 524f, 527  
 Lindman, H. 70, 204,  
     503, 524f, 527  
 Lindman, H. R. 94, 114,  
     181, 220  
 Lindquist, E. F. 57, 94,  
     102, 104, 110f, 128,  
     133, 144, 147, 149, 220  
 Linn, R. L. 136, 172,  
     199, 200, 220, 236, 334,  
     339, 343, 469  
 Lippman, L. G. 125, 220  
 Lissitz, R. W. 115, 220  
 LNR Research Group  
     563, 613  
 Löhr, H. J. 438, 462  
 Loftus, G. R. 149, 220  
 Lohnes, P. R. 94, 162,  
     201, 262, 456  
 Long, J. S. 378, 463  
 Loose, K. D. 422, 463  
 Lord, F. M. 40, 43,  
     107f, 130, 220, 240, 463  
 Lubin, A. 116, 147, 149,  
     218, 220  
 Lucas, P. A. 114, 208  
 Luce, R. D. 40, 217  
 Lübbcke, B. 110, 113,  
     210, 221  
 Lück, H. E. 66, 198  
 Lukas, G. 612  
 Lunney, G. H. 121, 221  
  
 MacCorquodale, K. 33,  
     221  
 Mace, A. E. 170f, 175,  
     221  
 Mackenzie, W. A. 107,  
     205  
 Mai, N. 395, 466  
 Makridakis, S. 266, 296,  
     463  
 Malinvaud, E. 463  
 Mandeville, G. K. 139f,  
     201, 214  
 Mann, H. B. 93, 221  
 Marascuilo, L. A. 90f,  
     93, 95f, 124, 143, 145,  
     219, 220, 221, 222, 285,  
     287, 289, 292, 462  
 Maritz, J. S. 524, 527  
 Marks, M. R. 89, 221  
 Markus, G. B. 385, 463  
 Marmor, M. 315, 463  
 Marmor, Y. S. 315, 463  
 Martens, H. R. 424,  
     439, 455  
 Martin, C. G. 112f, 221  
 Mason, S. J. 394, 423,  
     427, 438, 453  
 Massaro, D. W. 50, 221  
 Massey, F. J. Jr. 172,  
     174f, 202, 203  
 Massy, W. F. 395, 463  
 Matheson, D. W. 50,  
     221  
 Mauchly, J. W. 140, 221  
 May, R. B. 84, 127, 194,  
     221  
 McCain, L. J. 244, 248,  
     266f, 284, 442, 463  
 McCall, R. B. 20, 23, 94,  
     140, 143, 221, 295,  
     463  
 McCarthy, 551, 612  
 McCarthy, P. J. 414,  
     416, 454  
 McCleary, R. 244, 248,  
     259, 264, 266f, 277,  
     284, 311, 442, 463  
 McClelland, J. L. 420,  
     464  
 McDermott, J. 575,  
     577f, 612  
 McDonald, L. L. 181,  
     223  
 McDonald, R. P. 346f,  
     364, 464  
 McGill, W. J. 438, 464  
 McGinnis, R. 414, 464  
 McGrath, J. E. 50, 229  
 McGuigan, F. J. 50, 94,  
     153, 185, 222  
 McHugh, R. B. 90, 222  
 McKean, J. W. 117, 230  
 McLaughlin, D. 92, 118,  
     222, 235  
 McNeil, J. T. 296, 464  
 McNeil, K. A. 296, 464  
 McNemar, Q. 40, 222  
 McSweeney, A. J. 268,  
     464  
 McSweeney, M. 90f, 93,  
     95f, 124, 143, 221, 222  
 Meehl, P. E. 33, 35, 60,  
     68, 83, 189, 202, 221,  
     222  
 Mehta, J. S. 113, 222  
 Meier, F. 244, 303, 464  
 Melchinger, H. 328, 464  
 Melton, A. W. 84, 222,  
     511, 528  
 Mendenhall, W. 94,  
     105f, 222  
 Mendoza, J. L. 139ff,  
     216, 222, 228  
 Menges, G. 70ff, 79,  
     89f, 95, 104, 108f, 119,  
     222  
 Meredith, W. M. 92, 222  
 Mertens, W. 43, 222  
 Messick, S. 158, 163,  
     195  
 Metze, L. P. 201  
 Metz-Göckel, H. 239,  
     464  
 Metzler, P. 81, 84, 187,  
     217  
 Miles, J. A. 109, 230  
 Miles, R. F. 519, 526  
 Miller, A. D. 340, 464  
 Miller, G. A. 395, 464  
 Miller, G. R. 27, 222  
 Miller, H. D. 406f, 417,  
     456  
 Miller, J. J. 164, 222  
 Miller, R. G. 112, 118,  
     223  
 Miller, R. G. Jr. 120,  
     122, 124f, 223  
 Milliken, G. A. 181, 223  
 Mises, R. von 477, 528

- Möbus, C. 239, 242, 245, 247, 249, 305, 328, 384, 388, 450, 452, 464
- Montgomery, D. B. 395, 463
- Mood, A. M. 72, 223
- Moosbrugger, H. 94, 103, 105f, 137, 142, 223, 287, 464
- Morgenstern, O. 514, 528
- Moroney, M. J. 497, 528
- Morrison, D. E. 72, 223
- Morrison, D. F. 94, 140, 142, 223, 315, 347f, 359, 363, 365ff, 465
- Morrison, D. Y. 395, 463
- Moses, L. E. 56, 223
- Mosteller, F. 32, 54, 94, 102, 108, 117f, 174, 223
- Mourard, S. A. 164f, 213
- Mulholland, F. J. 438, 460
- Mundlak, Y. 370, 372, 375, 465
- Murphy, A. H. 509, 528, 529
- Murray, R. 125, 215
- Murray, J. R. 406, 417, 465
- Muthén, B. 343, 469
- Muthig, K. 28, 32, 94, 106, 108, 115, 144, 154, 164, 181, 211, 482, 490, 527
- Myers, J. L. 57, 94, 110, 123, 126, 129, 141, 223, 226
- Nagel, E. 477, 528
- Nagl, W. H. 338, 465
- Namoodiri, N. K. 48, 57, 59, 94, 107f, 114, 223, 465
- Narula, S. C. 182, 220, 223
- Nass, G. 84, 231
- Neave, H. R. 114, 223
- Nelson, C. R. 265, 465
- Nelson, P. L. 127, 223
- Nerlove, M. 348, 370f, 375, 454, 465
- Nesselrode, J. R. 239, 243, 454, 465
- Neumann, J. von 514, 528
- Newbold, P. 282, 459
- Newell, A. 406, 467, 532, 555ff, 563, 578, 598, 608, 612, 613
- Neyman, J. 71, 76, 78, 117, 189, 224, 449, 465, 476, 528
- Nicewander, W. A. 141, 163, 172, 201, 222, 224
- Nordholm, L. A. 165, 199
- Norman, D. A. 563, 613
- Novick, M. R. 43, 107f, 220, 524, 528
- Nowicki, S. 43, 198
- Nunnally, J. C. 83, 224
- Oakes, W. F. 86, 224
- Oakford, R. V. 56, 223
- O'Brien, R. G. 112, 113, 224
- Odeh, R. E. 127, 163, 174, 177, 224
- Österreich, R. 183, 189, 195
- Oliver, R. L. 58, 224
- Olkin, J. 165, 224
- Olson, C. L. 142, 163, 283, 224, 225
- O'Neill, R. 125, 225
- Opp, K.-D. 59, 225
- Oppenheim, P. 61, 221
- Orne, M. T. 43f, 225
- Orth, D. 39ff, 225
- Overall, J. E. 94, 112f, 130, 172, 178, 181, 225, 238, 524, 528
- Owen, D. B. 174, 225
- Owen, M. 299, 459
- Pack, D. J. 277, 465
- Padia, W. L. 455
- Page, E. B. 127, 225
- Parks, R. W. 374, 465
- Patry, J. L. 59, 225
- Patzig, G. 21, 23
- Pawlik, K. 239, 465
- Pearson, E. S. 54, 71, 78, 84, 110, 112f, 124, 174, 189, 224, 225, 226, 476, 528
- Pearson, K. 161, 226
- Peckham, P. D. 105ff, 130, 209, 296, 458
- Pedhazur, E. J. 5, 13, 94, 141, 162f, 180, 182, 215, 226, 296, 465
- Pertler, R. 171, 210
- Peterman, F. 67, 136, 149, 226, 234, 239, 333, 462, 465
- Peterson, N. L. 110ff, 211
- Petrinovich, L. F. 40, 120, 193, 226
- Pfanzagl, J. 40, 90, 113, 171, 185, 226
- Pfeiffer, P. E. 419, 465
- Phillips, L. D. 70, 226, 507, 524, 527
- Phillips, P. C. B. 430ff, 465
- Pierce, D. A. 266, 455
- Pillai, K. C. S. 142, 182, 226
- Pillemer, D. B. 30, 192, 226
- Pitman, E. J. G. 96, 226
- Please, N. W. 110, 226
- Plomp, T. 147, 150f, 154, 226
- Poltrock, S. E. 563, 612



- Poor, D. S. 138, 143, 226
- Popper, K. R. 16, 23, 30f, 226
- Porcia, E. 299, 460
- Posten, H. O. 110, 226
- Potthoff, R. F. 315, 318, 344, 465
- Powers, W. A. 207
- Pratt, J. W. 89, 95, 165, 208, 224, 226
- Preiser, S. 50, 227
- Price, J. M. 172, 224
- Priestley, M. D. 419, 465
- Probert, D. A. 112f, 208
- Przeworski, A. 422, 466
- Puri, M. L. 93, 95, 127, 227
- Quenouille, M. H. 56f, 94, 106, 214, 265, 466
- Raatz, U. 161, 166f, 193, 220
- Raghavarao, D. 57, 227
- Raiffa, H. 474, 514, 526, 527
- Rajalkashman, D. V. 454
- Ralston, M. L. 440, 466
- Ramsey, P. H. 125, 142, 227
- Ramseyer, G. C. 125, 227
- Rao, C. R. 315, 466
- Rapoport, A. 395, 421f, 466
- Rasch, D. 73, 94, 105, 108, 113, 136, 170f, 174, 187, 227
- Rashevsky, N. 422, 466
- Ray, W. S. 90, 227
- Redding, W. C. 59, 227
- Reichardt, Ch. S. 284, 466
- Remington, R. D. 32, 193
- Renn, H. 93, 95, 227, 239, 466
- Rennert, M. 136, 227
- Restle, F. 98, 227, 395, 438, 466
- Revenstorf, D. 136, 227, 239, 244, 266, 275, 296, 298, 300, 395, 408, 442, 466
- Richardson, L. F. 422, 467
- Richter, M. L. 30, 235
- Rieder, A. 103f, 193
- Roberts, R. R. J. 71, 84, 196
- Robinson, J. 91, 227
- Robinson, R. E. 40, 192
- Roden, A. H. 287, 458
- Rodger, R. S. 126, 175f, 227, 228
- Roed, J. C. 142, 210
- Röhr, M. 130, 228
- Rogan, J. C. 111, 113f, 124f, 139ff, 215, 216, 228
- Rogers, W. H. 117f, 192
- Romaniuk, J. G. 143, 228
- Ronning, G. 415, 467
- Rosenthal, R. 30, 43f, 62, 84, 201, 228, 229
- Roskam, E. E. 239ff, 334, 376, 467
- Rosnow, R. 43f, 229
- Rosnow, R. C. 62, 229
- Rosnow, R. L. 228
- Rost, J. 239ff, 467
- Rotton, J. 174, 229
- Rouanet, H. 139ff, 229
- Roy, S. N. 315, 318, 344, 465
- Rubin, D. B. 30, 43, 229
- Rubin, M. 296, 458
- Rudinger, G. 239, 467
- Rüppell, H. 70, 229, 524, 528
- Rützel, E. 18, 23, 70, 229, 524f, 528
- Rule, S. J. 121f, 229
- Rumelhart, D. E. 563, 613
- Runkel, P. J. 50, 229
- Ryan, T. A. 120ff, 229
- Rychener, M. D. 578, 613
- Sachdeva, D. 163, 229
- Sachs, L. 229
- Särndal, C. E. 164ff, 229
- Samiuddin, M. 112, 229
- Sanders, J. R. 105ff, 130, 209, 296, 458
- Sandler, H. M. 58, 213
- Saniga, E. M. 109, 230
- Sarris, V. 93, 230
- Savage, L. J. 70, 204, 478, 503, 524f, 527, 528
- Schach, S. 103f, 230
- Schäfer, T. 103f, 230
- Schäfer, W. D. 124, 202
- Schank, R. C. 563, 613, 614
- Scheffé, H. 90f, 104ff, 121, 125, 127, 141, 174, 230, 296, 350, 467
- Scheifley, V. M. 140, 230
- Schiefele, H. 239, 462
- Schlesselmann, J. J. 114, 230
- Schmidt, P. 59, 225
- Schmidt, W. H. 140, 230
- Schmidtke, A. 108, 230, 328, 454
- Schneeweiß, H. 521, 528
- Schoönemann, P. H. 174, 229
- Schrader, R. M. 117, 230
- Schubö, W. 82, 94, 106, 109, 182, 207
- Schucany, W. R. 118, 209
- Schuler, H. 36, 230

- Schulman, J. L. 84, 230  
 Schultz, R. F. Jr. 165, 203  
 Schulze, H.-H. 563, 598, 616  
 Schuss, Z. 430, 467  
 Schwartz, R. D. 66, 236  
 Schwarz, H. 73, 230  
 Schwarzer, R. 149, 230  
 Schweitzer, W. 418, 467  
 Scott, E. L. 449, 465  
 Scott, W. A. 43, 230  
 Searle, S. R. 105ff, 230, 295, 338f, 467  
 Seay, M. B. 30, 235  
 Sechrest, L. 66, 236  
 Seifert, H.-G. 72f, 77f, 82, 84f, 117, 210, 511, 527  
 Selg, H. 50, 56, 230  
 Sellnitz, C. 35, 201  
 Sen, P. H. 93, 95, 227  
 Shaffer, J. P. 89, 92, 125, 163, 218, 230, 231  
 Shampine, L. F. 440, 467  
 Shapiro, S. S. 109, 231  
 Shaughnessy, J. J. 45, 56, 235  
 Shaw, J. C. 532, 608, 613  
 Shenton, L. R. 110, 195  
 Sheridan, C. L. 50, 231  
 Shine, L. C. 287, 467  
 Shooter, M. 125, 216  
 Shuster, J. J. 124, 231  
 Siebel, W. 27, 231  
 Siegel, S. 93, 95, 143, 231  
 Silbey, V. 85, 199  
 Silverman, J. 43, 231  
 Silverstein, A. B. 93, 124, 231  
 Simon, H. A. 406, 421, 467  
 Simon, H. A. 532, 536, 538, 545, 555ff, 563, 598f, 608, 612, 613, 614  
 Simonton, D. K. 284, 315, 467  
 Singer, B. 93, 95, 231, 243, 406, 414, 437, 443, 467, 468  
 Skipper, J. K. Jr. 84, 231  
 Sligo, J. R. 143, 213  
 Slinde, J. A. 136, 220  
 Slovic, P. 524, 528  
 Slutzky, E. 251, 468  
 Smart, R. G. 71, 81, 84, 86, 231  
 Smith, H. 94, 203  
 Smith, J. E. R. 92, 114, 116, 118, 152, 231  
 Smith, J. L. 163, 231  
 Smith, R. A. 125, 231  
 Smitley, W. D. S. 95, 110, 114, 195  
 Shapper, K. 518, 526  
 Snedecor, G. W. 94, 106, 115, 232  
 Snell, J. L. 406, 461  
 Soares, G. A. D. 422, 466  
 Soderquist, D. R. 85, 163, 165, 168, 207, 232  
 Sodt, J. 85, 214  
 Sörbom, D. 376, 390ff, 433, 461, 468  
 Solomon, R. L. 58, 232  
 Som, R. K. 232  
 Spada, H. 239, 241f, 467  
 Spann, R. N. 394, 423, 427, 438, 453  
 Spector, P. E. 142, 232  
 Spiegel, D. K. 121, 144, 181, 214, 225  
 Spielman, S. 72, 232  
 Spilerman, S. 243, 406, 414ff, 437, 443, 467, 468  
 Spjøtvoll, E. 125, 232  
 Sprott, D. A. 130, 232  
 Srinivasan, R. 113, 222  
 Stäel von Holstein, C.-A.S. 509, 528  
 Stallone, F. 259, 461  
 Stanley, J. C. 22, 29, 32, 47f, 54, 58ff, 89, 111, 115f, 152f, 164, 181, 198, 209, 232, 284, 287, 456  
 Starbuck, W. H. 535, 611  
 Starmer, C. F. 92, 209  
 Stavig, G. R. 167f, 192, 232  
 Steffens, F. E. 125, 232  
 Steger, J. A. 40, 89, 232  
 Stegmüller, W. 28, 33f, 38, 61, 72, 232, 486, 499, 526, 529, 599f, 608f, 614  
 Stein, C. 175f, 233  
 Steinfatt, T. M. 40, 233  
 Steinhagen, K. 149, 230  
 Stellwagen, W. R. 127, 195  
 Sterling, T. D. 71, 81, 84, 233  
 Stevens, J. P. 142, 163, 183, 233  
 Stevens, S. S. 40, 233  
 Steyer, R. 181f, 233, 385, 468  
 Stilson, D. W. 72f, 77, 165, 233  
 Stoline, M. R. 232  
 Storm 396, 468  
 Straka, G. 239, 468  
 Streufert, S. 50, 206  
 Stuart, A. 72f, 89, 158, 215, 341f, 461  
 Student (W. S. Gosset) 233  
 Subkoviak, M. J. 172, 219, 233  
 Summers, G. F. 66, 233, 343, 469  
 Suppe, F. 28, 33, 61, 233  
 Suppes, P. 28, 38ff, 217, 233, 395, 457  
 Suran, B. G. 84, 230  
 Sutcliffe, J. P. 172, 233  
 Sutton, C. O. 153, 213  
 Swaminathan, H. 124,

- 234, 315, 318, 325, 346f,  
362, 264, 453, 464, 468  
Swanson, M. R. 124,  
198  
Tack, W. H. 243, 395,  
468  
Talwar, P. P. 112, 234  
Tamhane, A. C. 125,  
234  
Tang, P. C. 159f, 234  
Tatsuoka, M. M. 94,  
137, 234  
Taylor, C. J. 125, 220  
Tcheng, T.-K. 125, 227  
Teuter, K. 436, 438, 461  
Thalmaier, A. 451, 468  
Theil, H. 338, 340, 364f,  
468  
Thissen, D. 116f, 236  
Thomas, D. A. H. 125,  
234  
Thoresen, C. E. 287,  
292, 457, 468  
Thrall, R. M. 474, 526  
Tiao, G. C. 114, 196,  
282, 297, 304, 441, 455,  
496, 524, 526  
Tiku, M. L. 110, 118,  
234  
Timaeus, E. 43, 234  
Timm, N. H. 295f,  
315ff, 325, 328, 346f,  
352, 363, 366f, 369,  
384, 468  
Toothaker, L. E. 91,  
110f, 114, 125, 127,  
139ff, 204, 216, 222,  
223, 234  
Trabasso, T. 21, 22, 406,  
455  
Trachtman, J. N. 110,  
234  
Traxel, W. 50, 234  
Treiber, B. 85, 149, 234  
Treinies, G. 85, 234  
Trenkler, G. 93, 95f,  
109f, 198  
Trickett, W. H. 113, 235  
Tucker, L. R. 240, 468  
Tucker, R. K. 72, 85,  
199  
Tukey, J. W. 72, 94,  
107, 114, 116ff, 121,  
125, 127, 192, 202, 223,  
235  
Tuma, N. B. 243, 418,  
468, 469  
Turing, A. M. 596, 614  
Turner, E. D. 94, 165,  
207  
Tversky, A. 40, 98, 201,  
217  
Tyler, V. D. 285, 469  
Ueckert, H. 563, 565,  
575, 579, 591, 609, 614,  
615  
Underwood, B. J. 45,  
50, 56, 108, 235  
Unruh, W. R. 422, 463  
Upton, G. J. G. 92, 235  
Urry, V. W. 85, 230  
Ury, H. K. 125f, 235  
Valenzi, E. R. 85, 199  
Vatza, E. J. 30, 235  
Vaughan, G. M. 158,  
165, 235  
Veldman, D. J. 112, 238  
Venables, W. 165, 235  
Verreck, W. A. 149, 235  
Vigderhous, G. 244,  
268, 278f, 469  
Voevodsky, J. 420, 469  
Vogel, B. 408, 466  
Vogelbusch, W. 183,  
189, 195  
Wald, A. 71, 236, 494,  
497, 529  
Wald, W. 529  
Walker, A. M. 248, 262,  
469  
Wallace, T. D. 370, 375,  
469  
Wallasch, R. 239, 328,  
464  
Wallenstein, S. 141, 236  
Wallis, W. A. 93, 145,  
166, 217  
Walsh, J. E. 93, 236  
Walster, G. W. 172, 200  
Wang, Y. Y. 113, 164,  
236  
Wasserman, S. S. 243,  
469  
Waterman, D. A. 564,  
580, 615  
Watts, H. A. 440, 467  
Webb, E. J. 66, 236  
Weber, E. 71, 77, 236,  
498, 529  
Weber, S. J. 62, 236  
Webster, H. 240, 469  
Wegman, E. J. 118, 236  
Wegschneider, R. 395,  
466  
Weinberg, S. 172, 218  
Welch, B. L. 113, 235,  
236  
Wells, M. G. 86, 212  
Wender, K. F. 563, 598,  
616  
Wendt, D. 503, 506,  
509f, 512, 521, 526, 527,  
529  
Werts, C. E. 236, 334,  
339, 342f, 469  
Westermann, R. 29, 32,  
40ff, 69, 81, 97, 99,  
101f, 121f, 127, 144,  
189ff, 210, 236, 237  
Westmeyer, H. 28, 59,  
61, 63, 69, 209, 237  
Wetherill, G. B. 71, 95,  
125, 225, 237

- Wetzel, W. 136, 237  
 Wexler, K. 575, 616  
 Wheaton, B. 343, 469  
 Wheelwright, S. C. 266, 296, 463  
 Whitney, D. R. 93, 221, 237  
 Wickens, M. R. 430, 435, 465  
 Wiggins, A. D. 125f, 235  
 Wiggins, L. M. 243, 417, 469  
 Wike, E. L. 112f, 124, 199, 237  
 Wildt, A. R. 130, 237  
 Wiley, D. E. 240, 343f, 406, 417, 458, 465, 469  
 Wiley, J. A. 343, 469  
 Wilk, M. B. 104, 107, 109, 231, 237  
 Wilks, S. S. 348, 469  
 Willard, D. 299, 459  
 Williams, J. D. 125, 237  
 Willson, U. L. 245, 250, 252, 254, 266, 297, 299, 301, 303f, 459  
 Willson, V. L. 136, 209  
 Wilson, G. T. 470  
 Wilson, K. 141, 237  
 Wilson, R. S. 141, 237  
 Winer, B. J. 57, 94, 107, 113, 123f, 130, 133, 135, 138, 140f, 154, 160, 164, 174f, 182, 237, 319, 344f, 348, 357, 470  
 Winkler, H. B. 112f, 208  
 Winkler, R. L. 507, 509, 524, 528, 529  
 Winne, D. 175, 238  
 Wippich, W. 8, 23  
 Wishart, J. 163f, 238  
 Witte, E. H. 17, 23, 71f, 81, 90, 104, 238, 482, 490f, 500, 504, 511, 529  
 Wolfe, D. A. 93, 212  
 Wolfe, R. G. 406, 417, 465  
 Wooding, W. M. 116, 238  
 Woodward, A. J. 94, 238  
 Woodward, J. A. 112f, 130, 172, 178, 225, 238  
 Wormser, R. 37, 50, 144, 238  
 Wottawa, H. 43, 104, 106f, 238, 295, 422, 470  
 Würthner, K. 387, 460  
 Wundt, W. 60, 238  
 Wymer, C. R. 441, 454  
 Yates, F. 54, 56, 82, 84, 181, 205, 238  
 Young, A. A. 375, 460  
 Young, R. K. 112, 238  
 Yule, G. U. 248, 262, 335, 470  
 Zajonc, R. B. 385, 463  
 Zehnder, C. A. 165, 194  
 Zelen, M. 361, 462, 470  
 Zellner, A. 243, 410, 462  
 Ziegler, R. 3f, 19, 59, 213, 421, 470  
 Zimmermann, E. 26, 238  
 Zimmermann, P. 244, 470  
 Zimny, G. H. 50f, 56, 58, 238  
 Zinnes, J. L. 38ff, 233  
 Zurmühl, R. 380, 429, 470

# Sach-Register

- ARIMA-Modell 245ff
  - , allgemeines 253ff
  - ~ autoregressiver Prozesse 248ff
  - ~ integrierter Prozesse 246ff
  - ~ für moving average Prozesse 251ff
- Ausbalancieren 57
- Ausreißerwerte, Behandlung von 117f
- Autokorrelation 19, 248, 256ff
- Autoregressiver Prozeß 248ff, 259, 262, 348
- Bayes-Erwartungsmaximierungs-Kriterium 522f
- Bayes-Statistik 500ff, 507
- Computer-Simulation 20ff, 530ff
- Differenzwerte (s. Veränderungsmessung)
- Effekt, experimenteller 25, 81ff, 85ff, 157ff, 173f, 178f, 184, 186ff
  - , einfacher 146f, 149f
  - , Haupt- 146f, 151
  - , Interaktions- 147
- Effizienz, relative 95f
- Einfluß
  - , direkter 4ff
  - , indirekter 4ff
- Elimination 51, 131
- Entscheidungen, Kriterien für 519ff
  - Bayes-Erwartungsmaximierungs-Kriterium 522f
  - Laplace-Summen-Kriterium 521f
  - Minimax-Kriterium 521
  - Minimax-regret-Kriterium 521
  - Optimismus-Pessimismus-Kriterium 522
- Entscheidungsausgänge, Bewertung von 510ff
- Experiment 1ff, 24ff, 59f
  - , Sozialpsychologie des 43f
- Extremwerte s. Ausreißerwerte
- Falsifikationstheorie Poppers 16, 30f
- Falsifikation von Hypothesen 14ff, 30, 69f, 97, 122, 124, 188ff
- Falsifizierbarkeit (von Simulationsmodellen) 598ff
- Fehler erster Art ( $\alpha$ ) 14, 76ff, 84, 110f, 120f, 124f, 141, 173, 186f, 479
  - Kumulierung des 120f, 144
- Feiler zweiter Art ( $\beta$ ) 14, 78ff, 85, 102, 110f, 120, 122, 124f, 170, 173, 185ff, 480ff
  - Kumulierung des 122, 144
- Feldexperiment 59, 64, 66f, 190, 475
- Flußdiagramm 536ff
- Fourieranalyse 352ff
- Homogenisierung 131
- Hypothese, statistische 88ff, 477ff
  - , falsche 97ff
  - , gerichtete 88, 102
  - , nicht-parametrische 89, 166
  - , parametrische 89, 92, 158
  - , ungerichtete 88, 102
  - Verhältnis der - zur wissenschaftlichen Hypothese 14ff, 68f, 97f, 185
- Hypothesentesten, statistisches
  - Bayes-Verfahren 500ff
  - likelihood-quotienten-Test 495, 499
  - sequentielle Testverfahren 494ff
  - Signifikanztests 70ff, 185f, 484ff
- Intelligenz, künstliche 580ff
- Interaktion (statistische) 3, 64, 144ff
  - , disordinale 3, 64, 149ff
  - , ordinale 3, 150ff
- Intervallskala 39f, 94, 105, 114, 329, 473

- Kausalhypothese, Prüfung einer 1 ff, 27ff, 97, 188ff
- Konfundierung 42ff
- Konstanzhaltung 50f, 55, 131
- Konstrukt, hypothetisches 33
- Konstruktvalidierung 35
- Kontrollfaktor 55f, 128ff
- Kontrolltechniken 128ff
  - Ausbalancieren 57
  - eingenistete Faktoren 132f
  - Elimination 51, 131
  - Homogenisierung 131
  - Konstanzhaltung 50f, 55, 131
  - Kontrollfaktor 55f, 128ff
  - Kovarianzanalyse 130f
  - Parallelisierung 128ff
  - Randomisierung 2, 4, 52ff
  - wiederholte Messungen 56ff, 133ff
- Korrelationsforschung s. Korrelationsstudie
- Korrelationsstudie 59
- Kovarianzanalyse 13f, 105, 130f, 335f, 338, 376ff
  - von Zeitreihen von Querschnitten 376ff
- Kovarianzmodell (bei zeitbezogenen Daten) 338ff
- Laplace-Summen-Kriterium 521f
- lateinisches Quadrat 57
- LISREL-Modell 376ff, 423, 433ff, 449f
- Lord'sches Paradoxon 335ff
- Markoff-Modelle 395ff
- MaxKonMin-Strategie 144
- Messung 40f, 472ff
- Minimax-Kriterium 521
- Modell
  - , allgemeines lineares 103ff
  - , analytisches 533f, 550
  - , ARIMA- 245ff
  - , deterministisches 533, 535, 549
  - , dynamisches 18ff, 533, 549
  - , Entscheidungs- 533
  - , Erkundungs- 533, 535
  - , indeterministisches 533
  - , Kovarianz 338ff
  - , LISREL- 376ff, 423, 433ff, 449f
  - , Markoff- 395ff
  - , nicht-numerisches 550f
  - , probabilistisches 533f
  - , qualitatives 533f
  - , quantitatives 533f
  - , Regressions- 338ff
  - , Simulations- s. Simulationsmodell
  - , Strukturgleichungs- (zur Analyse von Veränderungen) 375ff
- Modellbildung 588ff
  - Abbildbarkeitsproblem (Repräsentationsproblem) 590f
  - Bedeutsamkeitsproblem (Testbarkeitsproblem) 590f
  - Eindeutigkeitsproblem (Transformierbarkeitsproblem) 590f
- Nominalskala 38, 92, 473
- Ordinalskala 38ff, 93, 114
- Panelanalyse 419ff, 443ff
- Parallelisierung 128ff, 143
- Parameterschätzung 505ff
  - Maximum-likelihood-Schätzung 506
  - Prinzip der kleinsten Quadrate 505f
  - Prinzip der robusten Schätzung 501ff, 507
- Pfadanalyse 5ff, 375
- Pfadkoeffizient 7ff
- Präzision (eines Experiments) 127ff, 185, 189
- Programmieren
  - , modulares 554f
  - bottom-up-programming 554f
  - top-down-programming 554
  - , nicht numerisches 550f
  - Listenverarbeitung 551ff
- Programmiersprache
  - Logo 538ff, 562, 565ff, 582f, 605f
  - LISP 551ff
- p-Werte (Signifikanztesten) 83f
  - , Fehlinterpretation der 84
- Quasi-Experiment 59, 67, 284
  - Zeitreihendesigns (s.a. Zeitreihendesign) 315ff
  - Zeitreihenexperimente (s.a. Zeitreihenexperiment) 284ff

## Querschnittsanalyse

„Pooling“ der  $\sim$  mit Zeitreihenanalyse  
370ff

Randomisierung 2, 4, 50, 52, 54, 59f,  
137, 287

Randomisierungstest 90f, 119, 287  
- bei N = 1-Experimenten 287ff

Reaktionsstile 43

Reliabilität 343

Replikation, konzeptuelle 35f

„response sets“ s. Reaktionsstile

Signifikanz, praktische (Effektstärke) 5,  
150f, 157f, 188ff, 490f

Signifikanztest 70ff, 185f, 484ff

-, einseitiger 88

F-Test 137ff, 172f, 295, 299, 302

Güteeigenschaften des - 484ff

multiple Mittelwertsvergleiche 123ff

nicht-parametrische Verfahren (s.a.

Randomisierungstest) 90ff

Robustheit der parametrischen Tests  
110ff

- bei N = 1-Experimenten 287ff

- bei N = 1-Zeitreihendesigns  
316ff

- bei wiederholten Messungen 137ff

- bei Zeitreihendesigns 316ff

- bei Zeitreihenexperimenten 287ff

t-Test 123ff, 172f, 295, 303

Varianzanalysen zur Analyse von  
Trends 344ff

-, zweiseitiger 88, 190

Simulationsmodell 538, 550, 554ff, 574,  
580, 587ff, 595ff

-, Prüfung des 595ff  
experimentalpsychologische

Prüfung 598f

Protokoll-Trace-Vergleich 597f

Turing-Test 596f

Validierung des 587ff

Skalenniveau 38ff, 91, 105, 114, 473f

Intervallskala 39f, 94, 105, 114, 473

Nominalskala 38, 92, 473

Ordinalskala 38ff, 93, 114, 473

Verhältnisskala 473f

Skalierung 472ff

Standardfehler 74, 127f, 185, 488

Stichprobenumfang, Bestimmung des  
14, 170ff, 486ff

$\sim$  bei multivariater Regressionsanalyse  
182f

$\sim$  bei multivariater Varianzanalyse  
182f

$\sim$  bei nominalen Daten 183

$\sim$  bei ordinalen Daten 183f

$\sim$  bei univariater Regressionsanalyse  
174ff

$\sim$  bei univariater Varianzanalyse  
174ff

Störfaktor 29, 34ff, 42, 46ff, 50, 52,  
54ff, 64f, 97ff

System, rekursives 3ff

Teststärke 79, 81, 85, 170, 173, 185f,  
480ff

Transformation, nichtlineare (von Zu-  
fallsvariablen) 114f

Validität 24f, 29ff, 155, 186, 286f

-, externe 29, 285

-, interne 2, 24, 29, 46ff, 67, 287

-, Populations- 24, 60ff, 149, 156,  
186

-, ökologische 286

-, Situations- 1, 24, 29, 63ff, 149,  
156

-, statistische 14ff, 25, 29, 67ff, 97ff,  
143f, 155f

-, Variablen- 24, 29, 33ff, 143

Variable

-, abhängige 1ff, 26, 338, 376

-, endogene 3

-, exogene 3, 312

-, implizite 5

-, latente 376

-, Stör- s.a. Störfaktor 1f

-, unabhängige 1ff, 26, 334, 338ff,  
376, 415

Varianzanalyse (zur Analyse von Trends)  
344ff

Veränderungsanalyse 239ff

-, systemtheoretische Sicht der 241f

Veränderungsmessung 136, 239ff  
 ~ mit Hilfe von Differenzwerten  
 328ff  
 Verhältnisskala 473f

Wachstumskurvenanalyse 344ff  
 ~ als Strukturvergleichsmodell 381ff  
 Wahrscheinlichkeit  
 -, Hypothesen- 493, 500ff, 507ff,  
 524f  
 übergangs- 410f, 415  
 Wahrscheinlichkeitsbegriff 477f  
 -, frequentistischer 477f  
 -, subjektivistischer 477f

Zeitreihenanalyse 243ff  
 -, multivariate 279ff, 419ff  
 Arima-Modell 281f  
 Transfermodelle 283f  
 „Pooling“ der ~ mit Querschnittsana-  
 lyse 370ff  
 -, univariate 243ff  
 Arima-Modelle (s.a. Arima Model-  
 le) 245ff  
 Transferfunktionsmodelle 268ff  
 Zeitreihendesign 315ff  
 multivariates - bei mehreren Grup-  
 pen 327f  
 univariates - bei einer Gruppe 315ff  
 Zeitreihenexperiment 284ff  
 N = 1-Experiment 285ff  
 Zufallsstichprobe 63, 73, 91, 119f